



**UNIL** | Université de Lausanne

Faculté de biologie  
et de médecine

Département d'Ecologie et Evolution

# DEVELOPMENTAL AND ANATOMICAL CONSTRAINTS ON VERTEBRATE GENOME EVOLUTION

Thèse de doctorat ès sciences de la vie (PhD)

Présentée à la Faculté de Biologie et de Médecine de l'Université de Lausanne par

JULIEN ROUX

Ingénieur en Biosciences, filière Bioinformatique et Modélisation, de

l'Institut National des Sciences Appliquées de Lyon, France

JURY

Prof. Mehdi Tafti, Président

Prof. Marc Robinson-Rechavi, Directeur de thèse

Prof. Jérôme Goudet, Co-directeur de thèse

Prof. Sven Bergmann, Expert

Prof. Henrik Kaessmann, Expert

Prof. Jianzhi George Zhang, Expert

Lausanne, 2010

# Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

Président	Monsieur Prof.	Mehdi Tafti
Directeur de thèse	Monsieur Prof.	Marc Robinson-Rechavi
Co-directeur de thèse	Monsieur Prof.	Jérôme Goudet
Experts	Monsieur Prof.	Henrik Kaessmann
	Monsieur Prof.	Sven Bergmann
	Monsieur Prof.	Jianzhi George Zhang

le Conseil de Faculté autorise l'impression de la thèse de

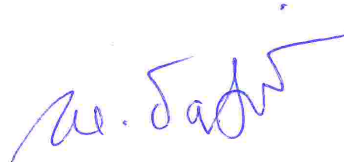
**Monsieur Julien Roux**

Ingénieur en biosciences de INSA, Lyon, France

intitulée

**DEVELOPMENTAL AND ANATOMICAL CONSTRAINTS  
ON VERTEBRATE GENOME EVOLUTION**

Lausanne, le 25 juin 2010



pour Le Doyen  
de la Faculté de Biologie et de Médecine

Prof. Mehdi Tafti

# Table of contents

---

Table of contents.....	i
Résumé de thèse.....	vii
Abstract.....	ix
Remerciements.....	xi
Introduction .....	1
Bgee: a tool to study the evolution of gene expression .....	2
State of the art .....	2
Presentation of Bgee.....	3
The need for (new) ontologies .....	3
Data integration pipeline .....	5
Data curation.....	5
Improvements .....	7
Microarray normalization .....	7
In situ hybridizations.....	7
Over-expression.....	8
Non expression.....	10
miRNAs .....	11
Contact with the community and users.....	11
Use of homology and related concepts into Bgee .....	12
Why is formalizing homology interesting beyond the needs of Bgee? .....	14
An ontology as a bioinformatics framework to represent homology-related concepts.....	15
How to deal with levels of organization? .....	18
Implementation into Bgee.....	19
Conclusion .....	20
Bgee: conclusion.....	21
The role of anatomy and development in the evolution of animal genomes and transcriptomes.....	22
Developmental Constraints on Vertebrate Genome Evolution .....	31
Molecular Signaling in Zebrafish Development and the Vertebrate Phylotypic Period .....	33

Expression in the nervous system drives retention after whole-genome duplication in vertebrates.....	34
Bibliographic references.....	36
1 Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species .....	45
Abstract .....	45
1.1 Introduction.....	46
1.2 Designing Homology Relationships between Anatomical Ontologies by an Ontology Alignment Approach.....	48
1.3 Mapping of the Developmental Ontologies .....	49
1.4 Integrating Heterogeneous Data on Anatomical and Developmental Ontologies...	50
Mapping Expression Data to Ontologies .....	50
Statistical Analyses.....	51
1.5 Database and Web-Interface of Bgee .....	52
1.6 Conclusions .....	53
1.7 Acknowledgements .....	53
1.8 References .....	54
2 An ontology to clarify homology-related concepts.....	57
Abstract .....	57
2.1 The problem: the concept of homology is divided by specialized usage.....	58
2.2 Towards a solution: an ontology of homology related terms .....	58
Similarity as root .....	59
Working definitions of homology .....	59
Multiple inheritance.....	59
Availability .....	60
2.3 Concluding remarks .....	60
2.4 Acknowledgments.....	60
2.5 References .....	63
3 Developmental Constraints on Vertebrate Genome Evolution.....	65
Abstract .....	65
3.1 Introduction.....	66
3.2 Results .....	67
Constraints on gene loss-of-function in zebrafish .....	67



Constraints on gene loss-of-function in mouse.....	69
Constraints on gene duplication.....	71
Constraints on gene sequence.....	73
Gene ontology characterization.....	74
3.3 Discussion.....	74
3.4 Materials and Methods.....	78
Microarray data.....	78
Significance of trends in zebrafish development.....	79
Clustering of microarray data.....	79
Mouse EST data.....	80
Zebrafish In-situ data.....	81
Rate of protein evolution.....	81
Genotypes and phenotypes.....	82
Zebrafish mutants.....	82
Zebrafish morpholinos.....	82
Mouse knock-outs.....	82
Identification of duplicate genes.....	83
Gene Ontology Analysis.....	83
Tools.....	84
3.5 Acknowledgements.....	84
3.6 Supporting Information.....	84
3.7 References.....	86
4 Molecular signaling in zebrafish development and the vertebrate phylotypic period..	91
Abstract.....	91
4.1 Introduction.....	92
4.2 Results.....	94
Protein interconnectivity is highest in early development.....	94
Signal transduction is highest in the larva.....	96
miRNA expression increases progressively through development.....	98
Characteristics of genes expressed during different developmental periods.....	99
4.3 Discussion.....	104
4.4 Conclusion.....	106
4.5 Material & Methods.....	107

Microarray data and clustering.....	107
Protein-protein interactions .....	107
Signal transduction genes.....	108
Gene Ontology analysis.....	108
Phenotypes and localization of expression data .....	109
miRNAs targets and expression.....	109
Conservation of gene expression in mouse .....	110
4.6 Acknowledgments.....	111
4.7 Supporting Information.....	111
4.8 Tables.....	112
4.9 References .....	116
5 Expression in the nervous system drives retention after whole-genome duplication in vertebrates.....	119
Abstract .....	119
5.1 Introduction .....	120
5.2 Results .....	122
Fish-specific whole-genome duplication .....	122
Bias in retention or evolution after duplication? .....	125
Vertebrate whole-genome duplications.....	125
Recent species-specific duplications .....	126
Number of isoforms .....	126
Nervous system expression and rate of sequence evolution .....	127
Explaining duplicate retention bias .....	128
Are duplicates slowly evolving genes?.....	129
5.3 Discussion .....	131
5.4 Material and Methods.....	134
Mouse expression data .....	134
Zebrafish expression data .....	134
Identification of duplicate genes .....	134
Ontology enrichment analyses.....	135
List of nervous system anatomical structures.....	135
List of anatomical structures from other systems.....	136
Number of isoforms .....	137

Rate of sequence evolution .....	137
Akashi's test.....	137
Mouse phenotypes.....	137
5.5 Acknowledgements .....	138
5.6 Tables.....	138
5.7 Supplementary tables.....	142
5.8 References .....	142
Outlook .....	147
Appendices .....	151
Appendix 1 .....	151
Appendix 2 .....	162
Appendix 3 .....	163
Appendix 4 .....	164
Appendix 5 .....	165



# Résumé de thèse

---

*Julien Roux*

*Contraintes Développementales et Anatomiques sur l'Evolution des Génomes Vertébrés*

*Département d'Ecologie et Evolution, UNIL*

Pendant ma thèse de doctorat, j'ai utilisé des espèces modèles, comme la souris et le poisson-zèbre, pour étudier les facteurs qui affectent l'évolution des gènes et leur expression. Plus précisément, j'ai montré que l'anatomie et le développement sont des facteurs clés à prendre en compte, car ils influencent la vitesse d'évolution de la séquence des gènes, l'impact sur eux de mutations (i.e. la délétion du gène est-elle létale ?), et leur tendance à se dupliquer. Où et quand il est exprimé impose à un gène certaines contraintes ou au contraire lui donne des opportunités d'évoluer. J'ai pu comparer ces tendances aux modèles classiques d'évolution de la morphologie, que l'on pensait auparavant refléter directement les contraintes s'appliquant sur le génome. Nous avons montré que les contraintes entre ces deux niveaux d'organisation ne peuvent pas être transférées simplement : il n'y a pas de lien direct entre la conservation du génotype et celle de phénotypes comme la morphologie.

Ce travail a été possible grâce au développement d'outils bioinformatiques. Notamment, j'ai travaillé sur le développement de la base de données Bgee, qui a pour but de comparer l'expression des gènes entre différentes espèces de manière automatique et à large échelle. Cela implique une formalisation de l'anatomie, du développement et de concepts liés à l'homologie grâce à l'utilisation d'ontologies. Une intégration cohérente de données d'expression hétérogènes (puces à ADN, marqueurs de séquence exprimée, hybridations *in situ*) a aussi été nécessaire. Cette base de données est mise à jour régulièrement et disponible librement. Elle devrait contribuer à étendre les possibilités de comparaison de l'expression des gènes entre espèces pour des études d'évo-devo (évolution du développement) et de génomique.



# Abstract

---

*Julien Roux*

*Developmental and Anatomical Constraints on Vertebrate Genome Evolution*

*Department of Ecology and Evolution, UNIL*

During my PhD, I used model species of vertebrates, such as mouse and zebrafish, to study factors affecting the evolution of genes and their expression. More precisely I have shown that anatomy and development are key factors to take into account, influencing the rate of gene sequence evolution, the impact of mutations (i.e. is the deletion of a gene lethal?), and the propensity of a gene to duplicate. Where and when genes are expressed imposes constraints, or on the contrary leaves them some opportunity to evolve. We analyzed these patterns in relation to classical models of morphological evolution in vertebrates, which were previously thought to directly reflect constraints on the genomes. We showed that the patterns of evolution at these two levels of organization do not translate smoothly: there is no direct link between the conservation of genotype and phenotypes such as morphology.

This work was made possible by the development of bioinformatics tools. Notably, I worked on the development of the database Bgee, which aims at comparing gene expression between different species in an automated and large-scale way. This involves the formalization of anatomy, development, and concepts related to homology, through the use of ontologies. A coherent integration of heterogeneous expression data (microarray, expressed sequence tags, *in situ* hybridizations) is also required. This database is regularly updated and freely available. It should contribute to extend the possibilities for comparison of gene expression between species in evo-devo and genomics studies.





# Remerciements

---

Je suis tout d'abord reconnaissant à mon jury de thèse pour la relecture du manuscrit et les commentaires intéressants lors de la soutenance privée du 1er Juin 2010. Plus particulièrement je tiens à remercier mon directeur de thèse Marc Robinson-Rechavi pour sa confiance et son soutien et pour m'avoir laissé une grande indépendance dans mes projets tout en laissant toujours sa porte grande ouverte pour discuter. Ces discussions ont souvent été une source d'idées, d'optimisme et de motivation.

Merci à mes collègues qui ont contribué à la bonne ambiance scientifique et humaine de l'équipe. Durant ces années j'ai directement travaillé avec plusieurs personnes dont le travail constitue une part importante de mon manuscrit de thèse. Spécialement un grand merci à Frédéric ; désolé pour les bugs dans le pipeline et le stress de leur découverte, et désolé pour tous les futurs à découvrir (j'espère pas trop nombreux). Je ne suis pas près d'oublier les discussions sans fin et les concours « un dîner presque parfait ». Merci à Aurélie, Fernando (salut mec !), Mar, Barbara, Sébastien, Anne et Walid. Merci aussi à tous les autres, y compris ceux qui ne sont pas restés longtemps : Vidhya, Romain, Alice, Antonia, Laurie, Patricia, Hannes, Grigoris, Estelle, Gilles, Yohan, Fred R.

Au Département d'Ecologie et d'Evolution pour l'ouverture d'esprit de ses membres. Même si la bioinformatique fait (faisait ?) peur à beaucoup d'entre vous, les occasions de discuter et d'exposer son travail sont toujours présentes. Merci aussi pour les distractions extra-professionnelles, les pauses café et les nombreux apéros. Il y aurait trop de monde à énumérer, les concernés se reconnaîtront ! Merci au personnel administratif et technique pour son efficacité et son aide quotidienne. Une dédicace spéciale à Yannick qui m'a contacté jusqu'au Danemark pour me motiver à venir faire ma thèse à Lausanne ... et merci pour tout le reste pendant ces années !

Au SIB, qui a contribué à mon financement pendant 3 ans, et à son fameux « PhD training network » pour les retraites enrichissantes à Bâle, Vevey ou Zürich. A Jean, Diana, Aitana, Armand, Yannick, Fred, Barbara, Daniel, Thomas, Charles, Luca, Antoine, Pascal pour les bons moments passés à ces occasions et ailleurs.

Au comité de l'ADAS (l'association des doctorants) pour l'organisation du D.Day chaque année et de manière générale pour la défense des doctorants. C'est important !

A mes potes de l'INSA, de l'UNIL ou d'ailleurs, toujours motivés pour excursions, bouffes, regroupements, VFE+n, etc. Merci à ceux qui se sont déplacés pour ma soutenance publique, c'était un chouette week-end ! Merci aux Vernier-Chichoux qui étaient là aussi.

A ma famille et plus particulièrement à mes parents et ma sœur Mathilde. Vous avez beaucoup contribué à me donner la ténacité, la curiosité et l'équilibre nécessaires pour faire une thèse. Merci de m'avoir toujours supporté dans mes choix, et de m'avoir donné les moyens d'élargir mes horizons !

Enfin merci infiniment à Edith, la personne qui compte le plus pour moi. Merci de m'avoir suivi à Lausanne et d'avoir eu la patience de supporter un thésard au quotidien (pas facile). Tu as su me changer les idées et me soutenir quand j'en avais besoin. Merci pour ta bonne humeur et pour tous ces bons moments passés avec toi.

# Introduction

---

I started my PhD with in Marc Robinson-Rechavi's lab in September 2006. During those four years, my work was focused on the two major aspects of bioinformatics. The first one is methodological and is illustrated by the development of tools and frameworks allowing the management of complex biological data. The second aspect is oriented to biology, and more precisely evolutionary biology. I made use of these tools and other data analysis methods to answer questions on the evolution of gene expression in vertebrates.

*Most of the original contributions presented in this thesis are, or intend to be, the object of refereed publications in journals or conferences. When I am not first author of the study, my contribution is presented at the beginning of the relevant section. The bibliographic references are displayed at the end of each relevant section.*

## **Bgee: a tool to study the evolution of gene expression**

Comparing different species can help the study of the evolution of organisms. It can also be useful for the study of species on which it is hard or impossible to experiment directly (typically human). Finally, multi-species comparisons are widely used to improve signal in genomics studies: transcription factor binding sites enhancing gene expression are more likely to be present in conserved regions of the genome, that can be uncovered using multiple alignments (see for example [1]).

During my PhD, partly funded by the Swiss Institute of Bioinformatics (SIB)<sup>1</sup>, I have been implicated in the development of the database Bgee (dataBase for Gene Expression Evolution), designed for the comparison of the transcriptome of different species. Studying the evolution of gene expression is important because it is underlying the evolution of phenotype. The evo-devo community is for example interested in understanding how changes in gene expression can affect morphology [2]. The conservation of gene expression in several species is also likely to reflect functionally relevant constraints acting on organisms.

We want to allow users to perform analyses on a high-throughput scale, with automatically computed results. This task is challenging, as it requires a complex integration of data. Bgee is available at: <http://bgee.unil.ch/>.

### ***State of the art***

An overview of the literature reveals two major types of studies comparing multiple species transcriptomes. A first one uses small-scale datasets to perform in-depth analysis of restricted systems (e.g. [3,4]). No standard format is used to store the data coming from such studies, and their integration seems problematic.

Another type of studies used higher scale datasets (tens to hundreds of tissues) in closely related species, such as human and mouse. Notably, a dataset generated by Su and colleagues [5], composed of microarray data for 79 tissues in human and 61 in

---

<sup>1</sup> The major aim of the SIB is to provide services and resources to the scientific community (e.g. Swissprot and StringDB).

mouse (<http://biogps.gnf.org/>), is widely used in comparative studies (the article has been cited more than 1000 times). A comparison of two mammals is rather easy due to the limited morphological divergence of most anatomical structures. It is not so easy to compare more distant species on such a scale, when important evolutionary transitions led to big morphological changes. Finally most studies make the approximation that all tissues are independent. This can be problematic when some tissues of the dataset are substructures of other tissues (for example hypothalamus and whole brain).

Several databases have emerged at the same time as Bgee, trying to address similar problems. *4DXpress* (<http://4dx.embl.de/4DXpress/>) focuses on *in situ* hybridization data, including mouse, zebrafish, medaka and fly. However the direct comparison between anatomical structures of different species is not implemented and the data have not been updated for the last two years. *Compare* (<http://compare.ibdml.univ-mrs.fr/>) exhibits similar functionalities to 4DXpress, but mainly redirects to different species-specific databases. Finally *BodyMap* (<http://bodymap.jp/>) allows the comparison of expression between multiple species, but only based on EST data. All tissues are mapped onto the human as a reference, limiting the possible investigations, and criteria for mapping are unclear.

This overview shows that there is a lack of resources addressing the problem of large-scale comparison of transcriptomes. Bgee aims to fill this gap.

### ***Presentation of Bgee***

An article describing Bgee was accepted for the conference “Data Integration in Life Sciences” (DILS) and published in June 2008 [6]. It is included in chapter 1. My part of the work is described in the sections 3 and 4 of the article. Below are discussed more in detail some specific aspects linked to my work.

#### *The need for (new) ontologies*

Ontologies are formal representations of knowledge within a domain, including concepts and relationships between them. They create a conceptual framework that computers can understand and reason on. They are nowadays frequently used for the description of gene function (the Gene Ontology [7]), the integration of high throughput ecological and evolutionary data [8,9,10,11], and are essential for the development of ambitious large-scale projects [12].

Using an ontology describing all tissues of an organism and the relationships between them (Figure 1) allows a formal encoding of expression patterns. Such ontologies have been developed by experts for anatomy and development of major model species.

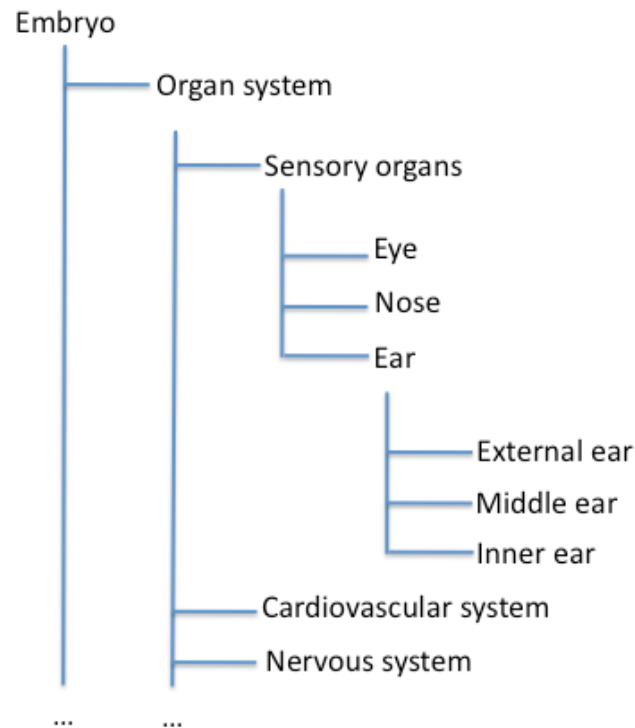


Figure 1: Part of the ontology describing the anatomy of mouse embryo [13].

The number of ontologies available for the community increases, but at a rather slow rate. The most recent ones include *Xenopus tropicalis* [14] and a common ontology for Teleost fishes [15]; an ontology describing *Platynereis* anatomy is in development. Ideally an anatomical and developmental ontology should be available for each animal species with a sequenced genome. The rapid rise of new sequenced model animal species (e.g. *Nasonia*, amphioxus, cichlid fishes [16]) will hopefully incite such efforts, but this is currently the major limiting factor for the integration of new species into Bgee.

To encourage such developments, the OBO foundry (Open Biomedical Ontologies)[17], a website gathering all biomedical ontologies, provides guidelines and principles for the creation of new ontologies. The consortium CARO (Common Reference Anatomical Ontology)[18] recently developed an ontology providing the basis for the development of new anatomical ontologies, with the aim that the resulting ontologies will be comparable and interoperable. Another approach is the development of a common

ontology for closely related species, such as teleost fishes for example, when morphologies are very similar.

### *Data integration pipeline*

When developing the pipeline for the integration of data into Bgee, the emphasis was on making it robust and adaptable. This is essential for the durability of the database. The field of biology is famous for the low persistence of resources available on internet [19], mainly due to the short term vision when developing them. Unlike in other fields, project specifications are rarely determined implementation begins. As a result, many databases are no longer updated and the exchange file formats or APIs (Application Programming Interface) are often modified. For these reasons, I tried to base our data extraction on reference databases (Ensembl [20], ZFIN [21], MGD [22], ArrayExpress [23]), which are more likely to have a long term view and guarantee regular updates. The code is easily adaptable when for example a new species is integrated.

Updates are regularly made on Bgee, aiming at following the rhythm of Ensembl releases. This required an optimization of the running time of the pipeline. Especially the insertion of probesets data for all Affymetrix chips analyzed into Bgee was optimized compared to an insertion using classical modules such as the Perl DBI. Without this optimization, this step alone would scale up to several weeks, considering the amount of data now available (more than a hundred million probesets inserted into Bgee in release 7).

I also paid special attention to the transmission of the project within the lab, with a comprehensive commenting of the pipeline code and a 'wiki' documentation (see appendix 1). For the latest update of Bgee (release 7), the pipeline was run jointly with Sébastien Moretti, bioinformatic programmer in the lab, who will be in charge of running it in the future.

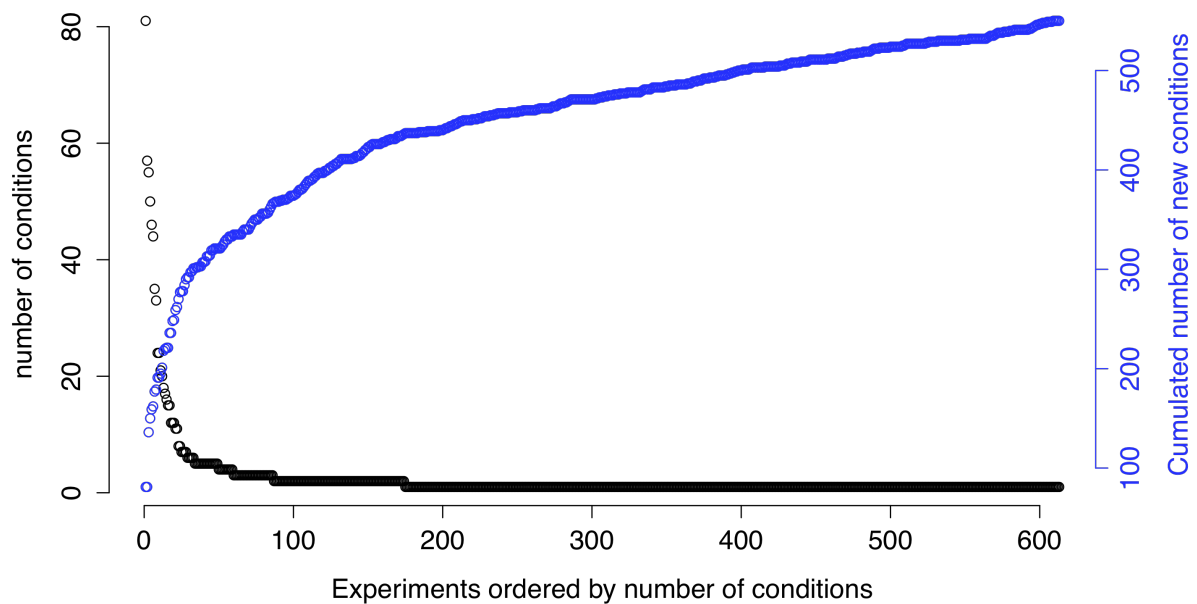
### *Data curation*

I am in charge of supervising data curation for Bgee with Frédéric Bastian (main developer of the Bgee application and database). An important effort was first put into the annotation of expression data. This mainly consists in annotating Affymetrix microarray experiments present in ArrayExpress [23] for zebrafish, fruit fly, mouse and human, to keep only microarray chips performed on untreated wild-type and healthy

samples (“normal” conditions), and annotate them with the corresponding term of the ontologies. Several master students have worked on this, and we have a permanent curator since March 2009; the number of annotated chips now exceeds 12000.

Nothing allows up to now the automation of this process. A recent development in ArrayExpress is the inclusion of an ontology of experimental factors (EFO)[10]. This will probably accelerate our curation process in the future.

We recently performed an analysis to check if the effort of microarray annotation for Bgee had not reached saturation. Figure 2 shows that annotating new experiments still brings new information into Bgee. If a plateau is reached one day, it may be interesting to focus on cleaning up the dataset and define new priorities for annotation.



*Figure 2: Overview of the annotation of microarray experiments in Bgee. A condition is defined as single organ at single developmental stage. We see that no plateau is yet reached, thus annotating new experiments, even small ones, still adds data for gene expression in new conditions.*

Of note our dataset made only of ‘normal’ conditions is of great interest for the community: we are notably in contact with the coordinators of ArrayExpress, who want to integrate this information in their database.



## ***Improvements***

Since the article was published, some new features were added to Bgee. They are discussed below.

### *Microarray normalization*

Most of the recent experiments that are deposited on ArrayExpress now provide raw data and we can renormalize them. Affymetrix microarrays are now routinely renormalized by the package gcRMA [24] of Bioconductor [25], which corrects probe signal for non-specific binding and probe sequence affinity.

It can be debated whether it is possible to make sense of present/absent summaries for gene expression with microarrays. We provide such summaries in Bgee because many experimental biologists do not consider gene expression as a continuum in practice, but rather as an ON/OFF pattern. This is supported by the observed bimodality of the intensity signal on many microarray experiments (see the bioconductor mailing list for example

<http://search.gmane.org/?query=bimodality&group=gmane.science.biology.informatics.conductor>). This observation can be made also with data from RNA-seq (Sarah Teichmann, personal communication), or fluorescence *in situ* hybridization (FISH) capable of detecting single mRNA molecules [26]. The method we use to detect presence or absence of expression was developed by Schuster et al. [27] and was shown to perform better than MAS5.

### *In situ hybridizations*

*In situ* hybridizations are widely used in developmental biology and evo-devo. They are precise and high-quality reports of the expression pattern of a given gene, at the level of fine anatomical structures. They do not require dissection of tissues, contrary to microarrays.

Several large-scale screens are currently ongoing in different organisms: the Thisse lab for zebrafish [28], Eurexpress for mouse embryos (<http://www.eurexpress.org/ee/>), Berkeley Drosophila Genome Project for early development of Drosophila [29].

We retrieve results from *in situ* hybridizations directly from the model organism databases (ZFIN [21], GXD [30], BDGP [29]). XenBase [31] recently implemented the management of *in situ* expression data for xenopus, and was added in the last release of

Bgee. The annotation of expression patterns from images on the anatomical ontologies is done directly by the curators of these databases.

Of note, the original development of anatomical ontologies was most often dictated by the use of high-precision annotations for *in situ* hybridization studies.

### *Over-expression*

It is of interest to extract “biologically pertinent” gene expression from microarray data, which might be more similar to the signal reported by *in situ* hybridizations. Besides present/absent information for a gene, it is interesting to have an idea about its specificity of expression. If a gene is expressed at a basal level in the body, but more highly expressed in a specific structure, it is probably relevant to focus on that structure. Experimentalists achieve this for *in situ* hybridizations by adapting manually the time of color development to get a good signal/noise ratio. The fixation step is done when the background noise – non-relevant expression – starts to increase.

For microarray data, this step cannot be done manually and a statistical analysis is needed to identify differential expression. I developed this framework with the help of Barbara Piasecka, PhD student in our lab. We use an ANOVA and the bioconductor package *Limma* [32,33], that implements a bayesian estimator of variance for expression values of probesets on the microarray. This is useful since microarray experiments usually contain few replicates, and since the variance of probesets depends strongly on the level of signal.

We kept for this analysis the experiments annotated in Bgee for which at least 3 “conditions” were studied on the same platform (type of array), and replicates were present for all conditions. A condition represents an organ at a developmental stage. For example adult brain and adult heart are two different conditions, as are embryonic brain and adult brain.

We next implemented a “multiple comparison to the mean” procedure (MCM) to identify the genes over-expressed in specific conditions. Each condition is contrasted to the global mean. Thus the contrasts performed are not independent ( $n-1$  independent contrasts can be performed for  $n$  conditions). To get rid of this problem, it is possible to compute simultaneous confidence intervals for these tests, using multivariate statistics (package *multcomp* in R). The downside of this approach is a greatly increased computational time. After discussions with Misha Kapushesky, who performs such

analyses for the ArrayExpress Atlas (re-analyses of high quality datasets of ArrayExpress; <http://www.ebi.ac.uk/gxa/>), we decided to keep the simple non-independent procedure because the results were globally not affected. It is anyway difficult to get rid of non-independence problems with expression data: neither genes nor tissues nor developmental stages are truly independent in one organism.

Figure 3 shows an example of such an analysis for the TMEM130 gene of a microarray experiment ([5]; 49 conditions analyzed). This protein is known to be part of the Golgi apparatus membrane, but its function is yet unknown (see <http://www.uniprot.org/uniprot/Q6NXM3>). The Bgee expression page is not very informative, with a total of 91 anatomical structures where the gene is expressed ([http://bgee.unil.ch/bgee/bgee?page=gene&action=summary&gene\\_id=ENSMUSG00000043388](http://bgee.unil.ch/bgee/bgee?page=gene&action=summary&gene_id=ENSMUSG00000043388)). The over-expression analysis however isolates 14 anatomical structures for this gene, 13 of them being substructures of the nervous system (the only non-nervous expression is in adult testes). This information is likely to be helpful in understanding the function of this protein.

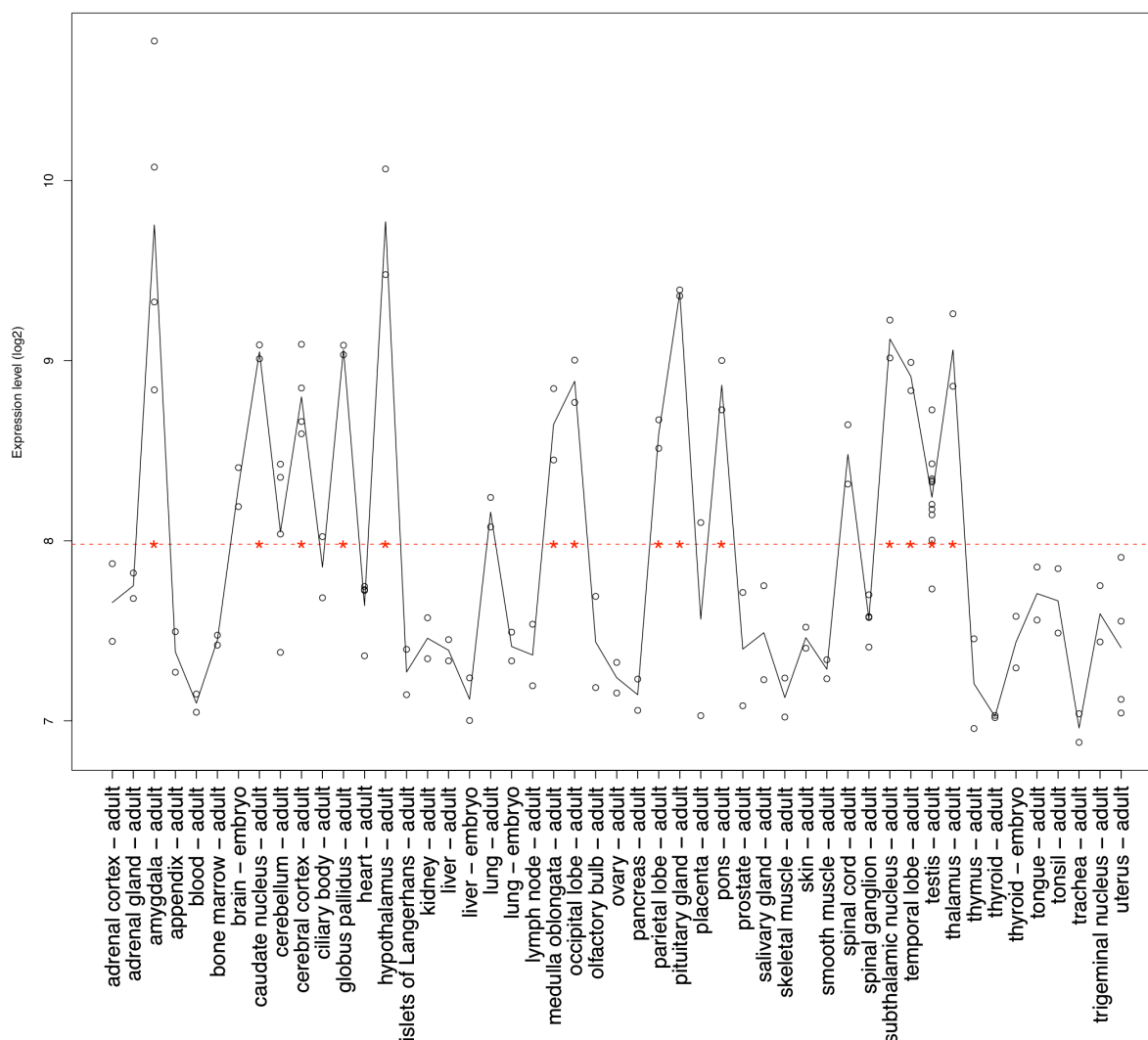


Figure 3: Expression profile of the probeset *gnf1h08859\_at* from the Su et al. dataset [5]. Dots represent expression measurements; black lines connect mean expression values for each condition; red dashed line is the mean value of expression for the probeset across all conditions analyzed; red stars mark conditions where significant over-expression is detected.

The visualization of differential expression data on the website of Bgee has yet to be implemented. They are currently only available in the MySQL database, or downloadable in flat files (<http://bgee.unil.ch/bgee/bgee?page=download>). For release 7 of Bgee, differential analyses were performed on 89 experiments and 4272 microarray chips (36% of Bgee microarray data).

### Non expression

It is currently possible to know whether a gene is expressed using Bgee. It is less clear however what can be said from the absence of expression data in Bgee. It can either be

due to a lack of data for the gene at a given organ or stage, or the gene may be truly not expressed (assuming that this has a biological meaning). We now explicitly store this information for microarrays (our analysis tells us if a given gene is above the background level on the chip or not), and for *in situ* hybridization data when it is reported by the curator – unfortunately this is rare.

### *miRNAs*

The pipeline of Bgee is based on Ensembl, which includes miRNA genes, but with very little descriptive information. This is in part because the mapping between Ensembl and species-specific databases is not complete. This blocks the automatic integration of expression data into Bgee for these genes (when retrieving *in situ* hybridization data for example). The phylogenies of miRNA genes are also not available in Ensembl, because these genes are too short to pass filters of automated pipelines of most databases.

The integration of miRNAs into Bgee was carried out by Mar Gonzàlez-Porta during a summer internship that I supervised. MiRNA families were extracted from miRBase [34], a specialized resource. A patch for cross-links allows the automatic retrieval of expression data from species-specific databases. A dataset of cloning profiles from smirnaDB [35] was also integrated for human, mouse, zebrafish and fly.

### ***Contact with the community and users***

It is important when developing a database to advertise it to potential users. We developed a brochure presenting Bgee, its interface and which questions it can address. The brochure is shown in appendix 2. Interactions with users are done through a mailing list: [bgee@isb-sib.ch](mailto:bgee@isb-sib.ch).

We also maintain regular contact with the bioinformatics community (ZFIN, Uniprot, neXtProt, ArrayExpress, 4DXpress, OBO foundry).

## Use of homology and related concepts into Bgee

Among future improvements planned for the database Bgee, a major concern is the use of more diverse concepts related to homology to compare anatomical structures. In the current release (release 7), Bgee allows the comparison of gene expression only between structures that derive from the same structure in the last common ancestor of the species considered. These structures are called 'historical homologs'. Historical homology is the most widespread definition for homology, and probably the best defined, but it is only a working definition. It does not accommodate all examples of recognized homology, and thus other definitions are needed [36]. This could be anecdotal, but it appears that some entire research fields, such as evo-devo, tend to favor alternative working definitions, because they fit better their interests [36,37]. As this community is a target for Bgee usage, it is important to develop a framework allowing the integration of different views of homology into Bgee.

Another lack of functionality of Bgee concerning anatomical structures comparisons appeared with the release 6 of the database (September 2009). In this release we integrated a new species: *Drosophila melanogaster*, as representative of arthropods and more broadly protostomes. In addition to its role as a major model organism, gene expression in the fruit fly can be used as an outgroup for vertebrates. But a major problem is that few structures are clearly homologous between *Drosophila* and vertebrate species. Demonstrations of homologies at this level of divergence require detailed investigations, as for example in the case of the central nervous system [38]. Currently most of the hypothesized homologies are still debated and it is not yet possible to come to a decision without in-depth studies. For example it is not clear if the ancestor of bilaterians was segmented, or if segmentation appeared independently in lineages leading to vertebrates and arthropods [39]. Finally many structures are not homologous at this taxonomic depth.

This does not prevent from wanting to compare the expression of genes in these structures, to learn about their characteristics and history. The most famous example of such cases is probably the eye: vertebrate and arthropods eyes are not homologous even if they are functionally equivalent. Surprisingly transcription factors implied in the

developmental cascade creating the eyes are conserved in both groups (including *Pax6* [40]). This result led scientists to reconsider the idea that eyes were the result of convergent evolution: it is now thought that these structures most probably evolved in parallel, originating from photoreceptor cells in their ancestor [41]. Another example is the developmental program of arthropod legs, implying the transcription factor *Dll*. It was co-opted for building numerous anatomical structures [42], including horns of beetles [43]. Legs and horns are neither homologous, nor analogous in the usual sense (i.e. nobody would have suggested homology), but this example teaches us that it is interesting to compare their expression to outline specific properties of patterning genes.

Thus it is important to develop for Bgee the functionalities allowing various types of comparisons: between homologous structures, including different working definitions of homology, analogous (or more precisely 'homoplastic') structures, functionally equivalent structures, and structures involving the expression of common developmental patterning genes ('homocratic'). A common denominator to all these relations is 'similarity': they all gather structures that resemble or are related to each other sufficiently to warrant a comparison. It would be possible to design Bgee to compare all anatomical structures showing some degree of 'similarity'. But this removes the ability to choose a level of granularity, if for example the user is interested in a specific type of homology.

Regarding this need, we could not rely on previous experiences or resources in the community. An overview of different projects dealing with homology between anatomical structures shows that some of them are restricted to strict historical homology because this is so far the most formalized concept (see for example <http://www.ohio.edu/phylocode/index.html>; this is also the approach in the current release of Bgee [6]). Some others chose a pragmatic approach and compare organs based on homonymy [5,44]. This may be appropriate when close species are compared (inside mammals for example), but comparisons of more distantly related species result in a mix of homology and homoplasy. Moreover similarity between homologous organ names is often not found when the structures diverged for a long time, if their function is not conserved, or simply because of different naming conventions in different communities (e.g. zoologists vs. medical doctors).

Faced with this situation, a first step has been to develop a bioinformatics framework able to deal with the complexity of concepts related to homology.

### ***Why is formalizing homology interesting beyond the needs of Bgee?***

Homology is an old concept, proposed by Owen in 1843 [45]. He defined homologs as “the same organ in different animals under every variety of form and function”. This concept survived, even though the use of vague terms such as “same” or “variety” might not fit our expectation of a specific definition. As opposed to many concepts created at the same period, homology proved to be central for evolutionary or comparative studies [46]. It is the relevant criterion to compare genes and organs, guaranteeing that the comparison makes sense in regard to evolutionary history: the properties of a structure in one organism are likely to be shared by the homologous structure in another organism. Homologies are also helpful to reconstruct phylogenetic history or to detect structures sharing common descent. Comparing the modifications that occurred since their last common ancestor can help to explain the adaptive modifications that these structures experienced.

One can grasp a feeling of the centrality of this concept by observing how it is a controversial topic in the community. Discussions, reviews, re-interpretation and re-definitions appear recurrently in the literature when new discoveries of biology challenge our view of homology.

Much of this confusion comes from the fact that we still do not know the underlying cause of homology. Homology, like *species*, is an *investigative kind* concept [37] (or *family resemblance* concept or *cluster* concept [47]). Brigandt describes “an investigative kind [as] a group of things that are assumed to belong together because they share a structural feature or mechanism that generates the characteristic features of the kind” [37]: some structures are thought to be homologous due to some interesting similarities that are perceived by scientists. However, the similarity is not what defines the homology. An underlying feature or process, yet unknown, is presumed to explain the observed similarities. Thus, the scientific search for the biological basis of homology is tightly linked to empirical work.

It appears that different research fields favor different operational definitions depending on their interests. Presently, it seems hard to find a universal definition of the concept of homology [46,48].



Thus the formalization of homology-related concepts needed for the development of Bgee can also be useful as a framework for future conceptual advance. Given the confusion resulting from numerous debates in the community, describing, clarifying and ordering all concepts in use, as well as the relations between them, may be an interesting way to go. Indeed this can provide a reference and a context for the proposal of new terms, contributing to avoid the repeated conflicts and redundancies observed so far. The current situation is too complicated to enforce "the only true meaning" of the homology-related concepts. This is probably because these concepts reflect directly the complexity of living organisms and their evolution [49].

With this aim, we did not limit our framework to concepts of potential use for Bgee in the short term (those used at the morphological level). In an effort to be exhaustive, we also included the numerous concepts used at other levels of organization, between genes for example. We limited our inventory to all terms with referenced use in the literature of the last decades. Gathering precise definitions otherwise dispersed among numerous articles and books may contribute to lower the hurdle to a good understanding of the concept of homology for biologists. Of note we did not consider the improper use of homology instead of similarity in molecules [50], as in reports of percentage of homology, or micro-homology at some positions.

### ***An ontology as a bioinformatics framework to represent homology-related concepts***

Ontologies provide a tool to organize complex related concepts. Orthogonality is respected here as no existing ontology already covers this specific field of knowledge [17]. We treated as synonyms the concepts which are redundant inside the field, and we drew relationships ('is\_a', 'part\_of') between concepts which are specific cases of each other. We also provided definitions and references for each described concept.

Concerning homology, some conceptually complex situations are easily represented with such an ontology. This is the case concerning the different working definitions of homology. It is able to reflect the multiple views on the definition of homology.

The 'historical homology' concept is one of these definitions, stating that two structures in two organisms are homologous if they derive from the same structure in the last

common ancestor [51,52]. One of the most classical examples of historical homology is probably the tetrapod limb, which is supported by a large fossil record that allowed the reconstruction of the successive evolutionary steps leading to the present picture [53,54]. This definition is widely used for taxonomic classification in cladistics: a homologous derived character shared by a group of taxa is called a synapomorphy and is a feature characterizing a monophyletic group or clade. For example the placenta is the innovation that distinguishes placental mammals (Eutheria). At the molecular level this definition proved to be particularly successful, because it is possible to reconstruct the history of genes and their families: by quantifying the similarity between sequences, we can evaluate their probability of common origin versus the probability of convergence. Comparing topologies of gene trees with species topology tells us if two sequences originated after a duplication, a speciation or a horizontal gene transfer. The specific terms for these cases are orthology, paralogy and xenology respectively [55,56,57]. Similarly a multiplicity of other sub-concepts has been created to describe specific evolutionary histories of genes (e.g. ohnology, equivalogy, interology, apparent orthology, in-paralogy). These can be ordered under the 'historical homology' concept in the ontology.

However at other levels, particularly at the morphological level, the historical definition does not help to recognize homologies in practice. Surprisingly, we still observe today that most of the criteria used for identifying homologies at that level have changed little since pre-Darwinian days and Owen's definition [37,48,58,59]. Similarity (or sameness) is at the basis of a homology statement. Looking at some structural parameters such as topology, connectivity of parts or developmental precursors is often enough to validate a homology hypothesis. Some concepts have appeared in the literature that detail or incorporate such knowledge into putative homology statements. For example 'homotopy' describes two homologous structures that share the same or similar relative positions. These concepts can be gathered in the ontology under the concept of 'structural homology' (or 'idealistic homology' [36,46]). Of note, it is traditionally considered in the community as a working definition, while strictly speaking it should be considered as an operational criterion to discover homologies in practice.

A third working definition addresses another problem of the historical homology definition: it does not fit all operational usages of the term homology. Some homologies are recognized that are not historical homologies. For example 'iterative homology' cannot be historical since it is a relation between structures of the same organism [51,60]. Some researchers tend to preserve the integrity of the historical homology definition and thus restrict the number of homology assessments. For them iterative homology is not a true homology. This solution, besides showing some circularity, does not fit some potential needs of biologists. This is the case for the field of evolutionary developmental biology (evo-devo), whose main questions are focused on how structures reappear *de novo* at each generation in different ontogenies. A different operational definition, 'biological homology', fitting evo-devo usage, was thus proposed by G. Wagner [36]. This concept is only defined at the morphological level. It is not focused on common ancestry, but rather is process-oriented and more mechanistic. Two structures are biological homologs if they are established and individualized similarly through development [36]. This includes repeated parts in the same organism (somites for example), as well as sexually differentiated parts of individuals of the same species (testis and ovaries for example). Together, these three definitions cover all legitimate uses of 'homology' in the modern literature.

It is interesting to note that the different working definitions are not disjoint, and most of the recognized homologies fulfill all of them [36,61]. Overall the cases of conflict are rather rare, and standard examples of homology or non-homology are the same for different research fields (the tetrapod limb for example).

To gather them, a common denominator is included in our ontology as a parent of the three different working definitions. It refers directly to some efforts in the literature to come up with a universal definition of homology, an umbrella or minimal approach including all cases of known homologies [62,63]. The required broadness imposes a rather vague definition. We chose to define it as 'inherited similarity', or similarity resulting from common evolutionary origin.

The use of this concept is also legitimate because it is the one that can be opposed to homoplasy (or analogy). Indeed the traditional view considers homology and homoplasy as disjoint concepts. Of note, 'biological homology' accepts a degree of ambiguity with homoplasy [36,64], because it does not focus on common ancestry. This ambiguity is

apparent in cases of latent homology, a form of parallelism between very similar structures occurring only within some members of a taxon and absent in the common ancestor: a lack of taxonomic resolution can easily lead to a hypothesis of biological homology. However it is likely that in most cases, a deeper investigation can lighten ambiguous cases and allow to attribute them to homology or homoplasy.

### ***How to deal with levels of organization?***

Twenty years ago, the unification of the field was foreseen, with the discovery that some patterning genes could be conserved over large evolutionary distances, from insects to vertebrates [65,66,67]. Could homologies result directly from the expression of these fundamental genes [68,69,70,71]? As attractive as it could be, this idea did not hold a long time before counter-examples were found. For example true homologies exist between structures sharing no conservation of expression patterns of underlying genes: the proteins of the vertebrate lens are unrelated and were co-opted [72]. Similarly the vulva of different species of nematodes are patterned by non-homologous pathways [73]. On the other hand, the conservation of expression of patterning genes alone is not sufficient to support an hypothesis of homology [37,42,59,61,68,69,74], as seen with the case of animal eyes.

Contrary to the expected clarification of what is homology, this has led to an increase in the degree of complexity of the concept. It became clear that homologies could refer to different levels of biological organization (anatomical structures, genes, developmental processes, behavior) and that they do not translate smoothly between these levels. Good practices recommend that homology statements should be made independently at each level or organization [61,68,74]. But some terminologies introduced in the literature show a mix of statements at different organization levels (patterning genes and anatomical structures most often)[41,75]. As an example, in the case of insect and vertebrate eyes, the patterning genes are homologous (*Eye* and *Pax6*), but the anatomical structures are not, they evolved in parallel. It is recommended to keep separate these two statements separate. However the term 'deep homology' has been proposed for such cases [41,76]. Instead, the term 'homocracy' can describe the relation between two structures that share homologous patterning genes, independently of any

homology assumption [42]. Therefore the recommended statement would be that vertebrate and insect eyes are homocratic but not homologous.

Such cases are easily represented in the ontology by multiple inheritance: deep homology is both a sub-concept of parallelism and of homocracy.

Another level that is frequently entangled in some terminologies related to homology is function [49,68,74]. Homologous structures often have the same function, because they derive from the same structure in the ancestor, but a statement of functional equivalence does not prove anything about homology: lungs and gills are used as respiratory organs in mammals and fish but are not homologous. Conversely some homologous structures have evolved different functions: the swim bladder in fish is homologous to lung in mammals ([77], p.210). Similarly tetrapod limbs are used for swimming, running, flying, climbing, etc. Still, we are often confronted with the term 'functional homology'. For example the term isoorthology, used to characterize orthologs having the same function [56], is in our ontology both a sub-class of orthology and of functional equivalence.

### ***Implementation into Bgee***

The developed ontology paves the way to the integration of homology-related concepts into Bgee. For his master project, Walid Gharib worked in our lab to set up the bases of this integration.

A first aspect of his work pinpointed that the integration requires more than a simple extension of the application currently used to run Bgee. Using only historical homology to compare organs, we currently group homologs into HOGs (Homologous Organs Groups). An idea is to create such groups for other concepts than homology. This works with those concepts that are transitive (e.g. homocracy, functional equivalence, biological homology). However some concepts in the ontology are not transitive. Homoplasy is one example. If the legs of insects are homoplastic compared to legs of horse and mouse, this does not imply that horse and mouse legs are homoplastic. Such relationships between organs thus have to be reported on an individual case basis, by creating pairwise relationships between organs, instead of grouping them.

The algorithms of the application have to be modified to take this into account. The retained solution involves the choice by the user of a reference species, on the basis of

which pairwise relationships and organs groups can be displayed. The user will also be able to choose a level of granularity, mirroring hierarchical levels of concepts in the ontology. Choosing ‘similarity’ (the root of the ontology) will display the totality of annotated relationships in the database, because all concepts in the ontology are sub-cases of similarity. On the opposite, a more specific research question of the user may for example require the choice of ‘serial homology’ (fifth level in the ontology), targeting a more restricted set of organs to query.

A final and important aspect to make this resource valuable, is then to provide exhaustive and high-quality annotations reflecting the state of the art in the specialized literature. The annotation effort of relationships between organs of four vertebrate species and *Drosophila* is currently ongoing, with the work of a curator, Aurélie Comte (until September 2009) then Anne Niknejad (since January 2010). I have been strongly involved in the coordination and supervision of their activity. Their work, integrated into the bioinformatics framework of Bgee, provides us with a unique dataset to perform large-scale studies of gene expression evolution.

## ***Conclusion***

To deal with homology and related concepts, our ontology presents an effort towards a solution to deal with the multiplicity and complexity of terminology in the literature, and the state of continuous debate of the field [46]. This bioinformatics resource is lasting, evolvable and re-usable by the community. It is deposited on the central repository of Biomedical ontologies, the OBO Foundry (<http://obofoundry.org/>; [17]).

A letter presenting the ontology was published in March 2010 [78]. It is included in chapter 2.

## **Bgee: conclusion**

We are just beginning to harvest the results of the work invested in Bgee. Its functionalities were already used for the evolutionary genomics studies described in chapters 3, 4 and 5.

As it gets mature, the evo-devo and evolutionary genomics communities should start to realize the increased possibilities to which this tool opens the door. One goal of Bgee is to reach a wide recognition in the next few years. It will be helped by future improvements to come, such as a addition of new expression data types (e.g. RNA-seq) and optimized way of treating them, or an extension of the possibilities of comparisons of anatomical structures of different species, allowing a rigorous handling of complex concepts related to homology. I hope that the expertise accumulated by the development of the solid basis of Bgee is an important asset in this ambitious task.

## **The role of anatomy and development in the evolution of animal genomes and transcriptomes**

The discovery of the structure of DNA [79] and of the genetic code [80] led to a dramatic development of the field of molecular biology, with immediate and profound repercussions on other fields of biology such as evolutionary biology. Countless examples are found in the literature of successful applications of molecular tools to study the evolution of organisms [81,82], some of them leading to paradigm shifts.

Probably one of the most striking examples of such shifts is the neutral theory of molecular evolution, formulated in the late 1960s by Kimura [83] and others, which changed the view of evolutionary biologists on the action of natural selection on genomes. While most of the mutations that reach fixation in a population were thought to do so as a result of adaptation and positive selection, Kimura's theory states that this is mainly due to a random process: genetic drift. Nearly neutral mutations are indeed invisible to natural selection if a critical size of reproducing individuals in the population is not reached. In more formal terms if the product of the effective population size ( $N_e$ ) with the effect of the mutation (selection coefficient  $s$ ) is much smaller than 1, stochastic genetic drift will overcome natural selection (see [84]). The power of natural selection will thus vary in different species, some species with large population size, as *Drosophila melanogaster*, experiencing stronger selection (in both directions, positive and negative selection) than others with small population size, such as human, whose effective population size is relatively small due to a population bottleneck during the out-of-Africa migration [85,86]. The neutral theory of evolution was largely validated with more molecular data in the following years [87]. Today the last points of the debate between neutralists and selectionists are being tackled by new experimental and theoretical advances [88,89]. But generally it is now accepted that non-adaptive mechanisms can have a major impact on the evolution of biological complexity, and that many trends of genomes evolution may not need adaptive explanations as additional hypotheses [82,90,91,92].

One factor that allowed these advances is an exponential increase of the amount of molecular data generated – for example nucleotidic sequences. The publication of the



first entire genome of a free-living organism, *Haemophilus influenzae* in 1995 [93] opened new avenues for *in-silico* analysis and bioinformatics applied to evolutionary questions. To date, more than 1500 complete or draft genomes are assembled, including around 100 animal species (<http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html>). Routine use of this amount of data is made for phylogenetics and comparative genomics, with for example applications to the prediction of protein functions using sequence homology.

Still, many questions are left open in evolutionary genomics. One of them is the importance of gene duplication. Gene duplication is a common mechanism, with a rate of duplication of the same order of magnitude as the rate of mutation per nucleotide site [94]. Some massive events of duplication of all genes of a genome at the same time have also been detected. For example the vertebrate lineage experienced three rounds of such whole-genome duplications, two of them in the ancestor of vertebrate, after the split with Cephalochordates and Tunicates [95]. A third one is specific to Teleost fishes [96]. We still do not understand fully why some genes duplicate more easily than others. Some debate is also ongoing on the consequences of duplication on phenotype or speciation abilities of the species that experienced them, but much remains largely speculative [97,98].

Another unanswered question is the role of non-coding DNA in the genome of animal species. In human for example more than 98% of the genome does not encode for a protein sequence. The mystery around this portion of the genome led some researchers to call it “junk” DNA [99]. It is now known to include essential functional elements, such as non-coding regulatory RNA genes [100] or *cis*-regulatory elements, as enhancers, playing essential roles in regulation of gene expression [101]. It has been found that some regions of non-coding DNA can be even more conserved in evolution than the protein-coding genes, suggesting a very strong action of purifying selection [102]. The large room for potential uncovered functionalities in the non-coding genome brings new hypotheses and theories. Recently in the field of evo-devo, Carroll and colleagues claimed that most of the mutations leading to the evolution of morphology in animals are located in *cis*-regulatory elements [2,103,104]. Due to the high level of pleiotropy of developmental patterning genes (such as homeobox domain containing genes), modular

changes in *cis*-acting elements would be the only way for genes to evolve new expression patterns. Indeed changes in the protein coding sequence of pleiotropic genes will lead to a multiplicity of effects and will be counter-selected. This would explain why the protein sequence of most patterning genes show a strong conservation (e.g. a mammal sequence can be developmentally functional when inserted in fruit fly). Several recent in-depth studies also support this theory [105,106,107,108,109]. However it is difficult to estimate if such case studies are anecdotic, or the first examples of a widespread phenomenon [110]. Some opponents of this theory also noticed that most of the reported examples are character losses, which may not involve the same evolutionary mechanisms as the (more interesting) character gains [111]. New approaches making use of large-scale genomics data have recently started to provide some clues about the relative proportion of *cis*- and *trans*-effects in evolution [112,113,114].

A last open question of evolutionary biology I will expand on here is related to the role of anatomy and development in the evolution of genomes. A few years ago, consistent results about evolutionary rates or patterns of proteins in bacteria and yeast seemed to indicate that a generalization was possible across the tree of life. "*Universals of protein evolution*" were thought to be responsible of most of the variance in protein features [115]. But more recently the extension of these rules to animal species (e.g. fruit fly, human, mouse, zebrafish) proved to be difficult. Many relationships between features and function of genes are influenced by their expression patterns in anatomy and development. During my PhD, I have conducted several analyses that focused on this influence. A graphical overview of the different protein features discussed in the context of my PhD thesis is presented in Figure 4.

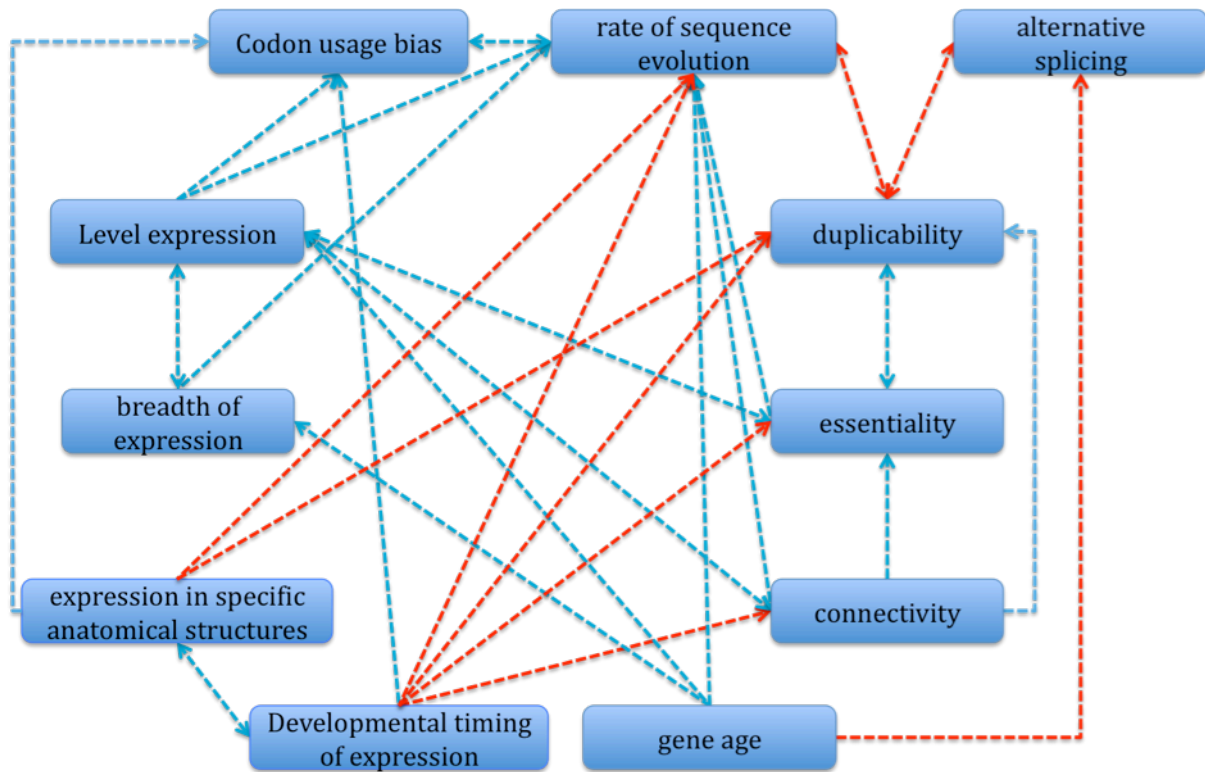
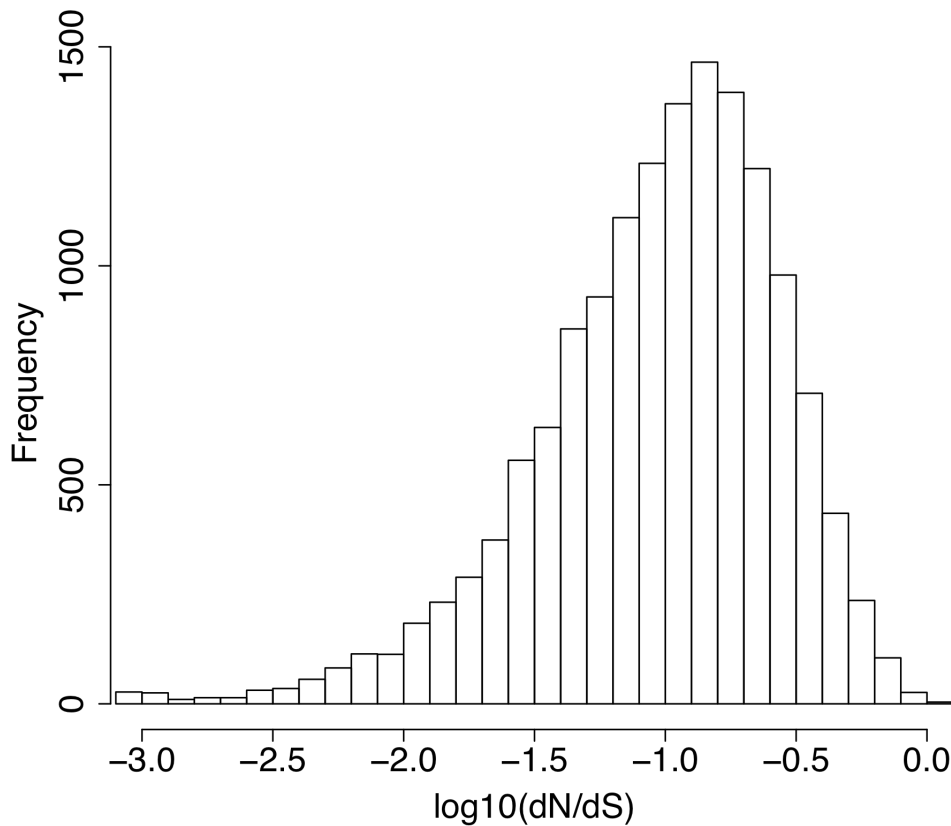


Figure 4: Different protein features and the relationships between them. The emphasis is on features related to gene expression and function. Structural features known to have an impact on protein evolution (e.g. length) are omitted from this graph. Arrows include causal relations, as well as possibly spurious ones reported in the literature. Red arrows represent relations that I could clarify or contribute to in my PhD work.

The feature that attracted the most attention over the past years is the rate of protein sequence evolution, probably because it is straightforward to measure on sequence data, with the ratio of  $d_N/d_S$  (rate of non-synonymous mutations over rate of synonymous mutations per site). This ratio can span several orders of magnitude in the same genome (see Figure 5), and this log-normal distribution is seen in a variety of species, ranging from bacteria to human [81]. Such a commonality among organisms showing a high diversity of phenotypes is unexplained [116].



*Figure 5: Distribution of the rates of protein sequence evolution the human genome. These were estimated using mouse one-to-one orthologs as outgroup The X-axis is in log-scale. Of note, almost all genes have a  $d_N/d_S$  lower than 1 (0 on log-scale) showing that they are under the action of purifying selection.*

An intuitive idea to explain differential rates of sequence evolution of proteins in a genome would be functional: more important proteins should be more highly conserved in evolution than less important ones, due to the action of purifying selection. In this perspective an association between protein rate of evolution and essentiality was tested, but it yielded a surprisingly low correlation (although statistically significant)[117,118]. Proteins essential for the fitness of an individual do not display the lowest rates of evolution.

A stronger and universal predictor of rate of protein evolution is rather the level of expression of the corresponding gene [119]: it has been suggested that a strong selective pressure is acting to prevent the consequences of protein misfolding. Indeed misfolding of a protein after its translation is first a waste for the cell, imposing the supplementary cost of destroying, and potentially decreasing fitness in conditions of intense growth.

Second the accumulation of misfolded proteins, prone to form hydrophobic aggregates that stick to the membranes, can be highly toxic [120]. This effect is logically stronger for highly expressed genes. Interestingly it has effects on both non-synonymous ( $d_N$ ) and synonymous substitution rates ( $d_S$ ) since an optimization of synonymous codon usage can improve the ability for a protein to fold properly. Because the rate of amino-acid misincorporation during the translation process is elevated, selection favors codons matching perfectly their corresponding tRNA, to improve accuracy of translation at functionally important sites.

Constraints on synonymous sites have also been found to be correlated to the level of expression due to selection for translation efficiency. Optimizing codon usage can also speed-up translation. This can be done by selecting codons matching the most abundant tRNA molecule for a given amino-acid [121]. This selection is so strong in some species of bacteria, that the measure of codon usage bias is often used directly as a measure of gene expression [115]. However selection for efficiency of translation can also be found in species such as *Drosophila* or mammals. Rapid progress is being made on that topic; for example a recent study uncovered a selection for codon co-occurrence in the translated mRNA sequence, allowing a faster recruitment of tRNAs [122]. Some recent studies also pinpointed the importance of optimizing the route of ribosomes on the mRNA molecule. Low codon efficiency at the beginning of the mRNA sequence is found in many organisms, while the last codons of the sequence are the most optimized [123]. This allows a reduction of ribosomal traffic jams, preventing segregation of ribosomes on highly expressed mRNA. The increased available pool of ribosomes in the cell, available for translation of other mRNAs, can directly affect the fitness of the organism [123,124]. The same scenario was also suggested after the observation that the secondary structure of the translated mRNA at ribosomal binding sites is fine-tuned to optimize the rate of ribosomal initiation [124].

Other significant factors acting on rate of protein sequence evolution include mutation and recombination rates, protein tertiary structure and its modularity, protein length [125], and protein-protein interactions [118]. Proteins that are more connected and more central in the network of protein-protein interactions tend to evolve slowly (although this is debated [126]) and are more essential [127,128]. The topology and modularity of networks was also shown to be important. For example the evolutionary

rates are different between proteins that interact with most or all of their partners simultaneously (“party” hubs) and those that interact with different partners at different times (“date” hubs)[129].

Finally a relationship was found between duplicability and the rate of protein sequence evolution: the genes that are more likely to be retained in duplicates are evolving slowly, as shown for duplications in yeast and nematode [130], as well as for the fish specific whole-genome duplication [131]. This result is counter-intuitive since duplicate genes have been shown to be less essential (probably because of some back-up of one duplicate by the other)[132,133], and thus should tend to evolve faster. This is indeed what we observed for genes that duplicated in human since the divergence with mouse (see chapter 6, Figure S23), consistently with some reports in literature [134,135]. The complex relation of duplicability with rate of sequence evolution may be due to other factors predisposing genes to be duplicated, such as the number of regulatory regions, the connectivity, or the level of expression [97].

A summary of the vast body of studies leaves us with a very complex network of relationships between protein features (Figure 4). Some of them are not fully understood because no mechanistic explanation could be found or tested rigorously to assert a causal explanation. This is for example the case for the link between duplicability and rate of protein sequence evolution. Such associations might be spurious, resulting only from correlations between covariables. The noisiness of some techniques (for example microarray data for level of expression, or yeast2hybrid for protein-protein interactions) can indeed lead to erroneous trends [136], and sometimes, methodological changes between studies led to opposite conclusions [115]. It is probable that technical advances will be useful to delineate more accurately some of these relationships, and uncover new ones. For example high-throughput sequencing allows a fantastic increase in accuracy for gene expression quantification (RNA-seq)[137]. Also, with the rapid development of proteomics and quantitative mass-spectrometry some studies recently suggested that it may be more pertinent to look at protein abundance than at transcript abundance, since the process of transcription itself appears to be very noisy, and appears indeed to be less conserved through evolution [138].

A coherent and global picture is thus yet to be integrated. Notably the generality of relationships has to be tested across different species. The vast majority of the reported studies have been conducted in baker's yeast (*Saccharomyces cerevisiae*). Because of an easy and convenient use in the laboratory, very advanced genetics techniques have been developed for this species, and the amount of data available is huge. Unfortunately, the limitations of a model unicellular eukaryote are being reached when testing the same relations in multicellular organisms, such as animals. A generalization is not straightforward because their complex anatomy and developmental processes adds a new layer of complexity and influences the evolution of protein-coding genes. Most correlations previously reported in unicellular organisms are not homogeneous across anatomy and development in vertebrates.

The relation between duplicability, protein connectivity and essentiality differs between mouse and yeast [139,140]. This was shown to be due to the confounding effect of the role of genes during development, affecting essentiality and duplicability [141]. The compensation between duplicates making them less essential, similarly to results in yeast, could be recovered in mouse by controlling both for functional role in development, and for protein network centrality [142].

In mammals, expression breadth (the number of tissues in which a gene is expressed) seems to explain evolutionary rates better than expression level [143,144]. In *Drosophila*, both the effect of maximum expression level and breadth of expression seem to have major roles [145].

Globally in vertebrates the effect of the level of expression on rates of sequence evolution is less strong than in other organisms. This is probably due to their lower effective population sizes and longer generation times. Because optimization of growth rate is unlikely to contribute to an increase of their fitness, little or no optimization of codon usage for translation efficiency is seen in protein sequences [144,146]. A small effect of selection can still be seen at some synonymous sites, helping to accurately translate functionally important residues, and contributing to robustness of proteins against misfolding [120].

Interestingly this trend is amplified for genes expressed in the nervous system, probably because toxicity of protein misfolding is likely to be more important in non-regenerating tissues. This hypothesis might explain also the slow rate of non synonymous mutations

in neural tissues [143]. Opposite to this pattern, expression in some tissues is correlated with a faster rate of sequence evolution. This is the case for example in testis, where divergence is probably led by sexual positive selection [143]. The contrast of the slow evolving brain-expressed genes with fast-evolving test-expressed genes is often seen in the literature (e.g. [147,148]).

In *Drosophila* too, expression in different anatomical structures probably has important consequences on the rates of protein evolution ([145], and also suggested by studies such as [149]), but few studies have focused on this aspect. However developmental timing of expression had received more attention. It is known to affect the rate of sequence evolution at non synonymous sites ( $d_N$ ) [150], but also at synonymous sites with codon bias variations across development [151,152]. This may be due to well-separated developmental periods (embryogenesis, larval stages, pupation) with probably very different selective forces acting on them. For example, the fruit fly larva experiences a drastic, almost exponential, increase of mass. Translation efficiency at this period directly affects the fitness of the individual, and is under strong selection. Genes expressed at this period are strongly biased in their codon usage. Genes expressed prior to this burst (in late embryogenesis) are selected for accuracy of translation, probably because the organism cannot afford the destruction of non-functional proteins during larval period (Roux and Petrov, in preparation). Developmental timing also seems to be an important factor to explain gene expression divergence between *Drosophila* species [153].

Thus it appears difficult to understand the evolution of protein-coding genes in animals without reference to their complex anatomy and development. The interplay between purifying selection and positive selection seems to be quite different between tissues and developmental stages.

During my PhD I focused on trying to complete and refine this picture, to understand the basic relations of features of protein evolution in the context of complex vertebrate organisms. The second part of this manuscript (chapters 3 to 6) describes several insights that emerged from my work. These chapters and their main implications are briefly discussed thereafter.



## ***Developmental Constraints on Vertebrate Genome Evolution***

Studies in nematodes, flies, and vertebrates have shown that the timing of expression of genes in development influences their evolutionary rate, some stages being constrained by purifying selection, while others show a higher tendency to be affected by positive selection [150,154,155,156]([reviewed in 157]).

In vertebrates however the picture was less clear [158,159]; no convincing difference could be shown between rates of protein evolution of proteins expressed at different time points during mouse development. This may be due to technical reasons, as for example the use of EST (Expressed Sequence Tags) counts, giving a very noisy estimate of gene expression levels. The lack of resolution during development may also hinder such studies: artificially dividing a continuous process into arbitrary broad developmental stages may yield a wrong picture. It may also be that true biological differences are present between protostomes and chordates, as their body plans and the way they are organized during development are very different.

Still, as embryonic development must proceed correctly for an individual to survive, vertebrates also should display some level of constraint preventing the accumulation through evolution of changes that have strong effects on the process of development. We examined whether changes that disrupt development too dramatically were indeed rare in evolution. The effect of mutations on coding sequences did not show a strong pattern but we could identify that the  $d_N/d_S$  was lower for genes expressed during embryonic development, compared to late stages and adult, in mouse and zebrafish.

We also investigated the effect of gene expression over vertebrate developmental time on two other features: the impact of mutation effects (i.e. is removal of the gene lethal?), and the propensity of the gene to remain in double copy after duplication. Duplicability and essentiality reflect constraints on gene dosage, and were shown to be related in yeast [160]. Both features are consistent, in both zebrafish and mouse, and indicate a strong effect of constraints in early development on the genome, constraints which are progressively weaker towards late development.

The implications of these results are manifold. First, we could clarify the pattern linking constraints across development with duplicability and essentiality. Selection preventing

gene dosage changes for genes expressed early in development is very similar to observation in yeast [160]. But this pattern did not hold later in development.

Second, while a hourglass model of morphological conservation in vertebrates has been observed since the 1990's (with a "phylotypic" period showing a maximum of conservation [161,162]), it has been more difficult to characterize the impact of such constraints on the genome. In this study we could show that the translation between those two levels of organization is not straightforward.

Third, contrary to observations that genes involved in developmental processes are preferentially retained after whole genome duplication (using analyses on Gene Ontology categories)[95,131], we show that genes expressed early in development are rather preferentially lost. This underlines that Gene Ontology annotations have to be taken with caution. The annotation of developmental processes is indeed made largely on genes implied in organogenesis and not very early development (Figure 6).

Fourth, the pattern that we uncover is not parallel to what is seen in *Drosophila*, where late embryogenesis seems to be under strong constraints [150](Alex Kalinka, personal communication). This suggests that different selective forces are experienced by vertebrates and arthropods during their development.

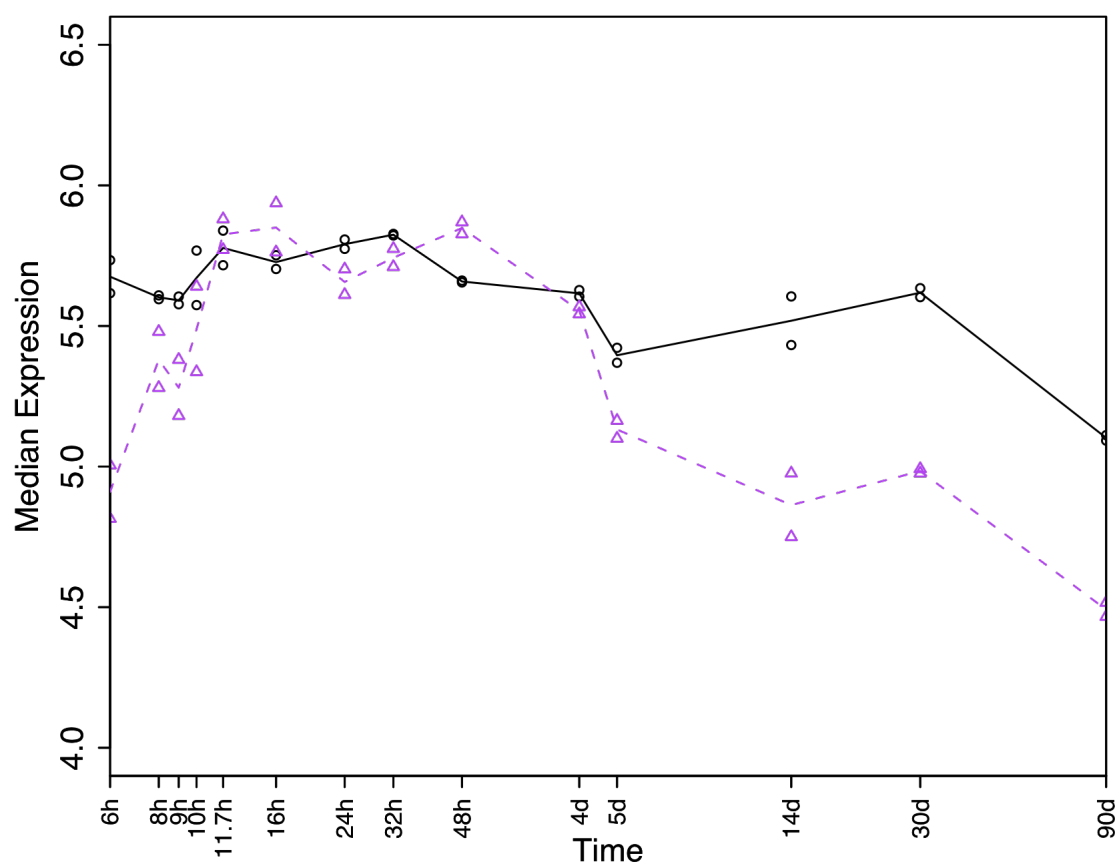


Figure 6: Median expression across zebrafish development of the genes annotated with the Gene Ontology category "developmental process" (in dashed purple line and triangles), compared to the median of all genes on the zebrafish Affymetrix microarray (in plain black line and circle). X-axis is in log-scale.

This study is presented in full in chapter 3. It was published in PLoS Genetics in December 2008 [163]. It was highlighted in Nature Reviews Genetics and Faculty of 1000 (see appendices 3 and 4).

### ***Molecular Signaling in Zebrafish Development and the Vertebrate Phylotypic Period***

This study is a follow-up and extension of the previous one, and focused on finding factors at the molecular level that could be responsible for an observed maximum of morphological conservation at the "phylotypic" period, around pharyngula [161,162,164,165]. More precisely we tested Evo-Devo claims on the role of embryonic signaling in conserved stages of vertebrate development. It has indeed been suggested

that high morphological conservation can be explained by an increased level of inductive interactions between organ primordia at this stage [162].

We could not support this hypothesis using molecular data in zebrafish. The network of protein interactions is denser early in development and this pattern decreases steadily though development, very consistently with the patterns of duplicability and essentiality across development. More central, more essential and less duplicable genes are seen early in development in vertebrates, but not at the phylotypic period.

Other molecular features, such as signal transduction cascades or miRNA activation could not be correlated with the hourglass pattern, implying that evolutionary or functional constraints at the molecular level do not explain morphological conservation of mid-development. High-level phenotypes such as morphology seem to be disconnected from patterns of genome evolution, and thus should not be used (at least in vertebrates) to infer selective constraints on genomes. Intuitive hypotheses involving concepts such as robustness or gene pleiotropy should for example be considered with great caution. Finally, modularity has been proposed to explain variation in morphological conservation across development. If some theoretical or small-scale examples of the benefits and consequences of modularity have been illustrated [166,167,168], it is usually hard to define clearly this concept and identify in real-life its influence on the evolution of organisms [169,170].

This study is presented in full in chapter 4. It was published in *Evolution and Development* in March 2010 [171].

### ***Expression in the nervous system drives retention after whole-genome duplication in vertebrates***

It is known that retention of duplicates after whole-genome duplication is not random. In fish for example, the genes retained in duplicate after the teleost-specific whole-genome duplication are evolving slowly, and belong to specific functional categories, such as development, signaling, behavior and regulation [98,131]. They are also genes expressed late in development (chapter 3).

We unraveled in this study a new bias regarding expression in anatomical structures. We find that genes expressed in structures of the nervous system are more likely to be retained in duplicate after such events. As genes expressed in the nervous system are

evolving slowly, the relation between duplicability and slow rate of evolution might thus be spurious. We show that it is partially the case, but that the relation is probably more complex.

Interestingly essential genes, which lead to lethality or sterility after knock-out, are not similarly over-expressed in nervous system structures. This indicates that neural tissues may be tolerant to gene dosage changes such as duplication or gene loss. It also underlines that essential genes should not be under strong selective pressure against protein misfolding, and this may explain why essentiality and rate of protein sequence evolution are only weakly correlated.

The high tolerance of neural tissues to duplication is interesting as it can lead to an increase of the repertoire of genes expressed in the nervous system. If this increase does not seem to be adaptive in the first place, it might be used later. Gene co-option was indeed hypothesized to have played an important role in the evolution of vertebrate nervous system [172].

This study is presented in full in chapter 5. It will be submitted soon.

## Bibliographic references

1. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15: 1034-1050.
2. Carroll SB (2005) *Endless Forms Most Beautiful: The New Science of Evo Devo and the Making of the Animal Kingdom*. W. W. Norton & Company. 350 p.
3. Tessmar-Raible K, Raible F, Christodoulou F, Guy K, Rembold M, et al. (2007) Conserved Sensory-Neurosecretory Cell Types in Annelid and Fish Forebrain: Insights into Hypothalamus Evolution. *Cell* 129: 1389-1400.
4. Escriva H, Bertrand S, Germain P, Robinson-Rechavi M, Umbhauer M, et al. (2006) Neofunctionalization in vertebrates: the example of retinoic acid receptors. *PLoS Genet* 2: e102.
5. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, et al. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA* 101: 6062-6067.
6. Bastian F, Parmentier G, Roux J, Moretti S, Laudet V, et al. (2008) Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species. *DILS: Data Integration in the Life Sciences*. pp. 124-131.
7. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25: 25-29.
8. Jones MB, Schildhauer MP, Reichman OJ, Bowers S (2006) The New Bioinformatics: Integrating Ecological Data from the Gene to the Biosphere. *Annu Rev Ecol Evol Syst* 37: 519-544.
9. Mabee PM, Ashburner M, Cronk Q, Gkoutos GV, Haendel M, et al. (2007) Phenotype ontologies: the bridge between genomics and evolution. *Trends Ecol Evol* 22: 345-350.
10. Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, et al. (2010) Modeling Sample Variables with an Experimental Factor Ontology. *Bioinformatics* 26: 1112-1118.
11. Washington NL, Haendel MA, Mungall CJ, Ashburner M, Westerfield M, et al. (2009) Linking Human Diseases to Animal Models Using Ontology-Based Phenotype Annotation. *PLoS Biol* 7: e1000247.
12. Genome 10K. Community of Scientists (2009) Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10 000 Vertebrate Species. *J Hered* 100: 659-674.
13. Aitken S (2005) Formalizing concepts of species, sex and developmental stage in anatomical ontologies. *Bioinformatics* 21: 2773-2779.
14. Segerdell E, Bowes JB, Pollet N, Vize PD (2008) An ontology for *Xenopus* anatomy and development. *BMC Dev Biol* 8: 92.
15. Dahdul WM, Lundberg JG, Midford PE, Balhoff JP, Lapp H, et al. (2010) The Teleost Anatomy Ontology: Anatomical Representation for the Genomics Age. *Syst Biol*: syq013.
16. Jenner RA, Wills MA (2007) The choice of model organisms in evo-devo. *Nat Rev Genet* 8: 311-314.
17. Smith B, Ashburner M, Rosse C, Bard J, Bug W, et al. (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 25: 1251-1255.

18. Haendel MA, Neuhaus F, Osumi-Sutherland D, Mabee PM, Mejino JLV, et al. (2007) CARO — The Common Anatomy Reference Ontology. *in* *Anatomy Ontologies for Bioinformatics: Principles and Practice*: Springer. pp. 327-350.
19. Wren JD (2008) URL decay in MEDLINE--a 4-year follow-up study. *Bioinformatics* 24: 1381-1385.
20. Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, et al. (2009) Ensembl 2009. *Nucl Acids Res* 37: D690-697.
21. Sprague J, Bayraktaroglu L, Clements D, Conlin T, Fashena D, et al. (2006) The Zebrafish Information Network: the zebrafish model organism database. *Nucl Acids Res* 34: D581-585.
22. Eppig JT, Bult CJ, Kadin JA, Richardson JE, Blake JA, et al. (2005) The Mouse Genome Database (MGD): from genes to mice--a community resource for mouse biology. *Nucl Acids Res* 33: D471-475.
23. Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, et al. (2009) ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression. *Nucl Acids Res* 37: D868-872.
24. Wu Z, Irizarry R, A., Gentleman R, Martinez-Murillo F, Spencer F (2004) A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association* 99: 909-917.
25. Gentleman R, Carey V, Bates D, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5: R80.
26. Raj A, Rifkin SA, Andersen E, van Oudenaarden A (2010) Variability in gene expression underlies incomplete penetrance. *Nature* 463: 913-918.
27. Schuster EF, Blanc E, Partridge L, Thornton JM (2007) Correcting for sequence biases in present/absent calls. *Genome Biol* 8: R125.
28. Thisse B, Heyer V, Lux A, Alunni V, Degraeve A, et al. (2004) Spatial and temporal expression of the zebrafish genome by large-scale in situ hybridization screening. *Methods Cell Biol* 77: 505-519.
29. Tomancak P, Beaton A, Weiszmamm R, Kwan E, Shu S, et al. (2002) Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. *Genome Biol* 3: research0088.0081-0014.
30. Smith CM, Finger JH, Hayamizu TF, McCright IJ, Eppig JT, et al. (2007) The mouse Gene Expression Database (GXD): 2007 update. *Nucl Acids Res* 35: D618-623.
31. Bowes JB, Snyder KA, Segerdell E, Jarabek CJ, Azam K, et al. (2010) Xenbase: gene expression and improved integration. *Nucl Acids Res* 38: D607-612.
32. Smyth GK (2005) Limma: linear models for microarray data. *in* *Bioinformatics and Computational Biology Solutions using R and Bioconductor*. New York: Springer. pp. 397-420.
33. Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 3: Article3.
34. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ (2008) miRBase: tools for microRNA genomics. *Nucl Acids Res* 36: D154-158.
35. Hausser J, Berninger P, Rodak C, Jantscher Y, Wirth S, et al. (2009) MirZ: an integrated microRNA expression atlas and target prediction resource. *Nucl Acids Res* 37: W266-272.
36. Wagner GP (1989) The Biological Homology Concept. *Annu Rev Ecol Syst* 20: 51.

37. Brigandt I (2003) Homology in comparative, molecular, and evolutionary developmental biology: The radiation of a concept. *J Exp Zool B Mol Dev Evol* 299B: 9-17.
38. Denes AS, Jekely G, Steinmetz PR, Raible F, Snyman H, et al. (2007) Molecular architecture of annelid nerve cord supports common origin of nervous system centralization in bilateria. *Cell* 129: 277-288.
39. De Robertis EM (2008) The molecular ancestry of segmentation mechanisms. *Proc Natl Acad Sci USA* 105: 16411-16412.
40. Halder G, Callaerts P, Gehring WJ (1995) Induction of ectopic eyes by targeted expression of the eyeless gene in *Drosophila*. *Science* 267: 1788-1792.
41. Shubin N, Tabin C, Carroll S (2009) Deep homology and the origins of evolutionary novelty. *Nature* 457: 818-823.
42. Nielsen C, Martinez P (2003) Patterns of gene expression: homology or homocracy? *Dev Genes Evol* 213: 149-154.
43. Moczek AP (2006) Integrating micro- and macroevolution of development through the study of horned beetles. *Heredity* 97: 168-178.
44. Haendel M, Gkoutos G, Lewis S, Mungall C (2009) Uberon: towards a comprehensive multi-species anatomy ontology. Available from Nature Precedings <http://dx.doi.org/10.1038/npre.2009.3592.1>.
45. Owen R (1843) Lectures on the comparative anatomy and physiology of the Invertebrate Animals: Delivered at the Royal College of Surgeons. London: Longman, Brown, Green, and Longmans.
46. Kleisner K (2007) The Formation of the Theory of Homology in Biological Sciences. *Acta Biotheoretica* 55: 317-340.
47. Pigliucci M (2003) Species as family resemblance concepts: The (dis-)solution of the species problem? *BioEssays* 25: 596-602.
48. Amundson R (2001) Homology and Homoplasy: A Philosophical Perspective. *Encyclopedia of Life Sciences*.
49. West-Eberhard MJ (2003) *Developmental Plasticity and Evolution*: Oxford University Press. 816 p.
50. Reeck GR, de Haën C, Teller DC, Doolittle RF, Fitch WM, et al. (1987) "Homology" in proteins and nucleic acids: A terminology muddle and a way out of it. *Cell* 50: 667-667.
51. Mayr E (1982) *The growth of biological thought: diversity, evolution, and inheritance*: Harvard University Press. 974 p.
52. Patterson C (1988) Homology in classical and molecular biology. *Mol Biol Evol* 5: 603-625.
53. Boisvert CA (2005) The pelvic fin and girdle of *Panderichthys* and the origin of tetrapod locomotion. *Nature* 438: 1145-1147.
54. Shubin NH, Daeschler EB, Jenkins FA (2006) The pectoral fin of *Tiktaalik roseae* and the origin of the tetrapod limb. *Nature* 440: 764-771.
55. Fitch WM (1970) Distinguishing Homologous from Analogous Proteins. *Systematic Zoology* 19: 99-113.
56. Fitch WM (2000) Homology: a personal view on some of the problems. *Trends Genet* 16: 227-231.
57. Koonin EV (2005) Orthologs, Paralogs and Evolutionary Genomics. *Annu Rev Genet* 39: 309-338.
58. Rutishauser R, Moline P (2005) Evo-devo and the search for homology ("sameness") in biological systems. *Theory in Biosciences* 124: 213-241.



59. Hall BK (1994) Homology: the hierarchical basis of comparative biology: Academic Press. 504 p.
60. de Beer GR (1971) Homology, an unsolved problem: Oxford University Press. 16 p.
61. Abouheif E (1997) Developmental genetics and homology: a hierarchical approach. *Trends Ecol Evol* 12: 405-408.
62. Scholtz G (2005) Homology and ontogeny: pattern and process in comparative developmental biology. *Theory in Biosciences* 124: 121-143.
63. van Valen LM (1982) Homology and causes. *Journal of Morphology* 173: 305-312.
64. Hall BK (2007) Homoplasy and homology: Dichotomy or continuum? *Journal of Human Evolution* 52: 473-479.
65. Gaunt SJ (1991) Expression patterns of mouse Hox genes: clues to an understanding of developmental and evolutionary strategies. *Bioessays* 13: 505-513.
66. Hanson I, Van Heyningen V (1995) Pax6: more than meets the eye. *Trends Genet* 11: 268-272.
67. Jones CM, Smith JC (1995) Inductive Signals: Revolving vertebrates. *Curr Biol* 5: 574-576.
68. Bolker JA, Rudolf AR (1996) Developmental genetics and traditional homology. *BioEssays* 18: 489-494.
69. Dickinson WJ (1995) Molecules and morphology: where's the homology? *Trends Genet* 11: 119-121.
70. Roth VL (1984) On homology. *Biological Journal of the Linnean Society* 22: 13-29.
71. Wagner GP (2007) The developmental genetics of homology. *Nat Rev Genet* 8: 473-479.
72. Piatigorsky J, Wistow G (1991) The recruitment of crystallins: new functions precede gene duplication. *Science* 252: 1078-1079.
73. Schlager B, Röseler W, Zheng M, Gutierrez A, Sommer RJ (2006) HAIRY-like Transcription Factors and the Evolution of the Nematode Vulva Equivalence Group. *Curr Biol* 16: 1386-1394.
74. Abouheif E, Akam M, Dickinson WJ, Holland PWH, Meyer A, et al. (1997) Homology and developmental genes. *Trends Genet* 13: 432-433.
75. Butler AB, William MS (2000) Defining sameness: historical, biological, and generative homology. *BioEssays* 22: 846-853.
76. Shubin N, Tabin C, Carroll S (1997) Fossils, genes and the evolution of animal limbs. *Nature* 388: 639-648.
77. Schmidt-Rhaesa A (2007) The evolution of organ systems: Oxford University Press. 385 p.
78. Roux J, Robinson-Rechavi M (2010) An ontology to clarify homology-related concepts. *Trends Genet* 26: 99-102.
79. Watson JD, Crick FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* 171: 737-738.
80. Nirenberg M, Leder P, Bernfield M, Brimacombe R, Trupin J, et al. (1965) RNA codewords and protein synthesis, VII. On the general nature of the RNA code. *Proc Natl Acad Sci USA* 53: 1161-1168.
81. Graur D, Li W-H (1999) Fundamentals of Molecular Evolution: Sinauer Associates Inc. 481 p.
82. Lynch M (2007) The Origins of Genome Architecture: Sinauer Associates Inc. 340 p.
83. Kimura M (1968) Evolutionary rate at the molecular level. *Nature* 217: 624-626.
84. Duret L (2008) Neutral theory: The null hypothesis of molecular evolution. *Nature Education* 1.

85. Eyre-Walker A, Keightley PD (2007) The distribution of fitness effects of new mutations. *Nat Rev Genet* 8: 610-618.
86. Hawks J, Hunley K, Lee S-H, Wolpoff M (2000) Population Bottlenecks and Pleistocene Human Evolution. *Mol Biol Evol* 17: 2-22.
87. Kimura M (1991) The neutral theory of molecular evolution: A review of recent evidence. *The Japanese Journal of Genetics* 66: 367-386.
88. Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, et al. (2009) Genome evolution and adaptation in a long-term experiment with *Escherichia coli*. *Nature* 461: 1243-1247.
89. Wagner A (2008) Neutralism and selectionism: a network-based reconciliation. *Nat Rev Genet* 9: 965-974.
90. Duret L, Galtier N (2009) Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. *Annu Rev Genomics Hum Genet* 10: 285-311.
91. Lusk RW, Eisen MB (2010) Evolutionary Mirages: Selection on Binding Site Composition Creates the Illusion of Conserved Grammars in *Drosophila* Enhancers. *PLoS Genet* 6: e1000829.
92. Duret L, Galtier N (2009) Comment on "Human-Specific Gain of Function in a Developmental Enhancer". *Science* 323: 714c.
93. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269: 496-512.
94. Lynch M, Conery JS (2000) The Evolutionary Fate and Consequences of Duplicate Genes. *Science* 290: 1151-1155.
95. Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, et al. (2008) The amphioxus genome and the evolution of the chordate karyotype. *Nature* 453: 1064-1071.
96. Jaillon O, Aury J-M, Brunet F, Petit J-L, Stange-Thomann N, et al. (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* 431: 946-957.
97. Sémon M, Wolfe KH (2007) Consequences of genome duplication. *Curr Opin Genet Dev* 17: 505-512.
98. Kassahn KS, Dang VT, Wilkins SJ, Perkins AC, Ragan MA (2009) Evolution of gene function and regulatory control after whole-genome duplication: Comparative analyses in vertebrates. *Genome Res* 19: 1404-1418.
99. Ohno S (1972) So much "junk" DNA in our genome. *Brookhaven symposia in biology* 23: 366-370.
100. Mattick JS (2004) RNA regulation: a new genetics? *Nat Rev Genet* 5: 316-323.
101. Visel A, Rubin EM, Pennacchio LA (2009) Genomic views of distant-acting enhancers. *Nature* 461: 199-205.
102. Woolfe A, Goodson M, Goode DK, Snell P, McEwen GK, et al. (2005) Highly Conserved Non-Coding Sequences Are Associated with Vertebrate Development. *PLoS Biol* 3: e7.
103. Carroll SB (2008) Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. *Cell* 134: 25-36.
104. Wray GA (2007) The evolutionary significance of cis-regulatory mutations. *Nat Rev Genet* 8: 206-216.
105. Chan YF, Marks ME, Jones FC, Villarreal G, Jr., Shapiro MD, et al. (2010) Adaptive Evolution of Pelvic Reduction in Sticklebacks by Recurrent Deletion of a *Pitx1* Enhancer. *Science* 327: 302-305.

106. Gompel N, Prud'homme B, Wittkopp PJ, Kassner VA, Carroll SB (2005) Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in *Drosophila*. *Nature* 433: 481-487.
107. Prud'homme B, Gompel N, Rokas A, Kassner VA, Williams TM, et al. (2006) Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. *Nature* 440: 1050-1053.
108. McGregor AP, Orgogozo V, Delon I, Zanet J, Srinivasan DG, et al. (2007) Morphological evolution through multiple cis-regulatory mutations at a single gene. *Nature* 448: 587-590.
109. Jeong S, Rebeiz M, Andolfatto P, Werner T, True J, et al. (2008) The Evolution of Gene Regulation Underlies a Morphological Difference between Two *Drosophila* Sister Species. *Cell* 132: 783-793.
110. Wagner GNP, Lynch VJ (2008) The gene regulatory logic of transcription factor evolution. *Trends Ecol Evol* 23: 377-385.
111. Hoekstra HE, Coyne JA (2007) The locus of evolution: evo devo and the genetics of adaptation. *Evolution* 61: 995-1016.
112. Tirosch I, Reikhav S, Levy AA, Barkai N (2009) A Yeast Hybrid Provides Insight into the Evolution of Gene Expression Regulation. *Science* 324: 659-662.
113. Fraser HB, Moses AM, Schadt EE (2010) Evidence for widespread adaptive evolution of gene expression in budding yeast. *Proc Natl Acad Sci USA* 107: 2977-2982.
114. Lemos B, Araripe LO, Fontanillas P, Hartl DL (2008) Dominance and the evolutionary accumulation of cis- and trans-effects on gene expression. *Proc Natl Acad Sci USA* 105: 14471-14476.
115. Rocha EP (2006) The quest for the universals of protein evolution. *Trends Genet* 22: 412-416.
116. Wolf YI, Novichkov PS, Karev GP, Koonin EV, Lipman DJ (2009) The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. *Proc Natl Acad Sci USA* 106: 7273-7280.
117. Herbeck JT, Wall DP (2005) Converging on a general model of protein evolution. *Trends Biotechnol* 23: 485-487.
118. Pal C, Papp B, Lercher MJ (2006) An integrated view of protein evolution. *Nat Rev Genet* 7: 337-348.
119. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH (2005) Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci USA* 102: 14338-14343.
120. Drummond DA, Wilke CO (2008) Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. *Cell* 134: 341-352.
121. Duret L (2002) Evolution of synonymous codon usage in metazoans. *Curr Opin Genet Dev* 12: 640-649.
122. Cannarozzi G, Schraudolph NN, Faty M, von Rohr P, Friberg MT, et al. (2010) A Role for Codon Order in Translation Dynamics. *Cell* 141: 355-367.
123. Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, et al. (2010) An Evolutionarily Conserved Mechanism for Controlling the Efficiency of Protein Translation. *Cell* 141: 344-354.
124. Kudla G, Murray AW, Tollervey D, Plotkin JB (2009) Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* 324: 255-258.
125. Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL (2005) Evolution of Proteins and Gene Expression Levels are Coupled in *Drosophila* and are Independently

- Associated with mRNA Abundance, Protein Length, and Number of Protein-Protein Interactions. *Mol Biol Evol* 22: 1345-1354.
126. Stumpf MPH, Kelly WP, Thorne T, Wiuf C (2007) Evolution at the system level: the natural history of protein interaction networks. *Trends Ecol Evol* 22: 366-373.
  127. Prachumwat A, Li W-H (2006) Protein Function, Connectivity, and Duplicability in Yeast. *Mol Biol Evol* 23: 30-39.
  128. He X, Zhang J (2006) Why Do Hubs Tend to Be Essential in Protein Networks? *PLoS Genet* 2: e88.
  129. Fraser HB (2005) Modularity and evolutionary constraint on proteins. *Nat Genet* 37: 351-352.
  130. Davis JC, Petrov DA (2004) Preferential duplication of conserved proteins in eukaryotic genomes. *PLoS Biol* 2: e55.
  131. Brunet FG, Crollius HR, Paris M, Aury JM, Gibert P, et al. (2006) Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol Biol Evol* 23: 1808-1816.
  132. Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, et al. (2003) Role of duplicate genes in genetic robustness against null mutations. *Nature* 421: 63-66.
  133. He X, Zhang J (2006) Higher Duplicability of Less Important Genes in Yeast Genomes. *Mol Biol Evol* 23: 144-151.
  134. Han MV, Demuth JP, McGrath CL, Casola C, Hahn MW (2009) Adaptive evolution of young gene duplicates in mammals. *Genome Res* 19: 859-867.
  135. Huminiecki L, Wolfe KH (2004) Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. *Genome Res* 14: 1870-1879.
  136. Drummond DA, Raval A, Wilke CO (2006) A single determinant dominates the rate of yeast protein evolution. *Mol Biol Evol* 23: 327-337.
  137. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10: 57-63.
  138. Schrimpf SP, Weiss M, Reiter L, Ahrens CH, Jovanovic M, et al. (2009) Comparative functional analysis of the *Caenorhabditis elegans* and *Drosophila melanogaster* proteomes. *PLoS Biol* 7: e48.
  139. Liang H, Li WH (2007) Gene essentiality, gene duplicability and protein connectivity in human and mouse. *Trends Genet* 23: 375-378.
  140. Liao BY, Zhang J (2007) Mouse duplicate genes are as essential as singletons. *Trends Genet* 23: 378-381.
  141. Makino T, Hokamp K, McLysaght A (2009) The complex relationship of gene duplication and essentiality. *Trends Genet* 25: 152-155
  142. Liang H, Li W-H (2009) Functional compensation by duplicated genes in mouse. *Trends Genet* 25: 441-442.
  143. Gu X, Su Z (2007) Tissue-driven hypothesis of genomic evolution and sequence-expression correlations. *Proc Natl Acad Sci USA* 104: 2779-2784.
  144. Duret L, Mouchiroud D (2000) Determinants of Substitution Rates in Mammalian Genes: Expression Pattern Affects Selection Intensity but Not Mutation Rate. *Mol Biol Evol* 17: 68-70.
  145. Larracunte AM, Sackton TB, Greenberg AJ, Wong A, Singh ND, et al. (2008) Evolution of protein-coding genes in *Drosophila*. *Trends Genet* 24: 114-123.
  146. Chamary JV, Parmley JL, Hurst LD (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet* 7: 98-108.

147. Xu Q, Modrek B, Lee C (2002) Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucl Acids Res* 30: 3754-3766.
148. Voolstra C, Tautz D, Farbrother P, Eichinger L, Harr B (2007) Contrasting evolution of expression differences in the testis between species and subspecies of the house mouse. *Genome Res* 17: 42-49.
149. Chintapalli VR, Wang J, Dow JAT (2007) Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease. *Nat Genet* 39: 715-720.
150. Davis JC, Brandman O, Petrov DA (2005) Protein evolution in the context of *Drosophila* development. *J Mol Evol* 60: 774-785.
151. Hense W, Anderson N, Hutter S, Stephan W, Parsch J, et al. (2010) Experimentally Increased Codon Bias in the *Drosophila* Adh Gene Leads to an Increase in Larval, but Not Adult, Alcohol Dehydrogenase Activity. *Genetics* 184: 547-555.
152. Vicario S, Mason CE, White KP, Powell JR (2008) Developmental stage and level of codon usage bias in *Drosophila*. *Mol Biol Evol* 25: 2269-2277.
153. Rifkin SA, Kim J, White KP (2003) Evolution of gene expression in the *Drosophila melanogaster* subgroup. *Nat Genet* 33: 138-144.
154. Artieri C, Haerty W, Singh R (2009) Ontogeny and phylogeny: molecular signatures of selection, constraint, and temporal pleiotropy in the development of *Drosophila*. *BMC Biol* 7: 42.
155. Castillo-Davis CI, Hartl DL (2002) Genome Evolution and Developmental Constraint in *Caenorhabditis elegans*. *Mol Biol Evol* 19: 728-735.
156. Yang J, Li WH (2004) Developmental constraint on gene duplicability in fruit flies and nematodes. *Gene* 340: 237-240.
157. Garfield D, Wray G (2009) Comparative embryology without a microscope: using genomic approaches to understand the evolution of development. *J Biol* 8: 65.
158. Irie N, Sehara-Fujisawa A (2007) The vertebrate phylotypic stage and an early bilaterian-related stage in mouse embryogenesis defined by genomic information. *BMC Biol* 5: 1.
159. Hazkani-Covo E, Wool D, Graur D (2005) In search of the vertebrate phylotypic stage: a molecular examination of the developmental hourglass model and von Baer's third law. *J Exp Zool B Mol Dev Evol* 304: 150-158.
160. Wapinski I, Pfeffer A, Friedman N, Regev A (2007) Natural history and evolutionary principles of gene duplication in fungi. *Nature* 449: 54-61.
161. Duboule D (1994) Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. *Dev Suppl*: 135-142.
162. Raff RA (1996) *The shape of life: genes, development, and the evolution of animal form*. Chicago; London: University of Chicago Press. 520 p.
163. Roux J, Robinson-Rechavi M (2008) Developmental Constraints on Vertebrate Genome Evolution. *PLoS Genet* 4: e1000311.
164. Haeckel E (1874) *Anthropogenie oder Entwicklungsgeschichte des Menschen* Leipzig: Engelmann. 732 p.
165. von Baer KE (1828) *Ueber Entwicklungsgeschichte der Thiere: Beobachtung und Reflexion*. Königsberg: Bornträger. 271 p.
166. von Dassow G, Meir E, Munro EM, Odell GM (2000) The segment polarity network is a robust developmental module. *Nature* 406: 188-192.
167. Klingenberg CP (2008) Morphological Integration and Developmental Modularity. *Annu Rev Ecol Evol Syst* 39: 115.

168. Wagner GP, Pavlicev M, Cheverud JM (2007) The road to modularity. *Nat Rev Genet* 8: 921-931.
169. Bininda-Emonds OR, Jeffery JE, Richardson MK (2003) Inverting the hourglass: quantitative evidence against the phylotypic stage in vertebrate development. *Proc Biol Sci* 270: 341-346.
170. Galis F, Metz JA (2001) Testing the vulnerability of the phylotypic stage: on modularity and evolutionary conservation. *J Exp Zool* 291: 195-204.
171. Comte A, Roux J, Robinson-Rechavi M (2010) Molecular signaling in zebrafish development and the vertebrate phylotypic period. *Evol Dev* 12: 144-156.
172. Holland LZ (2009) Chordate roots of the vertebrate nervous system: expanding the molecular toolkit. *Nat Rev Neurosci* 10: 736-746.

# 1 Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species

---

Frederic Bastian\*, Gilles Parmentier\*, Julien Roux, Sebastien Moretti, Vincent Laudet,  
Marc Robinson-Rechavi

*My major contribution to this project is the development of a pipeline for the integration of biological data into Bgee. In this article my part of the work is described in the sections 3 and 4.*

## Abstract

Gene expression patterns are a key feature in understanding gene function, notably in development. Comparing gene expression patterns between animals is a major step in the study of gene function as well as of animal evolution. It also provides a link between genes and phenotypes. Thus we have developed Bgee, a database designed to compare expression patterns between animals, by implementing ontologies describing anatomies and developmental stages of species, and then designing homology relationships between anatomies and comparison criteria between developmental stages. To define homology relationships between anatomical features we have developed the software Homolonto, which uses a modified ontology alignment approach to propose homology relationships between ontologies. Bgee then uses these aligned ontologies, onto which heterogeneous expression data types are mapped. These already include microarrays and ESTs. Bgee is available at <http://bgee.unil.ch/>

This article was published in Data Integration in Life Sciences (2008) Bairoch A, Cohen-Boulakia S, and Froidevaux C (Eds.) Lecture Notes in Computer Science 5109: 124-131.  
[doi: 10.1007/978-3-540-69828-9 12](https://doi.org/10.1007/978-3-540-69828-9_12)

---

\* Co first authors.

## 1.1 Introduction

Gene expression patterns (when and where a gene is expressed) are a key feature that underlies the development of organisms and phenotypes of individuals. They are an important aspect of the study of gene function. Moreover, the study of the evolution of developmental processes, often called “evo-devo”, has shown that the primary source of change in the evolution of phenotypes is changes in gene expression [1] rather than sequence.

Comparing gene expression patterns between animals is thus a major step in the study of gene function as well as of animal evolution, and also provides a link between genes and phenotypes.

In biological research, results obtained in different organisms are routinely compared. A comparative approach may be chosen for practical reasons because the organism of interest (humans, farm animals) may be less amenable to experimentation than more or less distant model species (as mouse, rat, zebrafish, or fruit fly).

Another reason is that components of gene expression may vary for no obvious reason [2]; this introduces the problem of distinguishing this signal from the noise caused both by random evolution and the inaccurate data measurements. Comparative study of gene expression in several species may contribute to this distinction. For example, comparing multiple samples from humans and rodents gave sufficient statistical evidence for a functionally relevant component of gene expression [3], and allowed for significant improvement in tumour characterisation [4].

Transcriptome data have also been compared among species to gain direct insight into evolutionary processes. For instance, yeast microarray data provided evidence for divergence of expression after genome duplication [5], and further studies have succeeded in extracting some evidence for the evolution of new gene functions after genome duplication in yeast and human lineages [6, 7]. A comparative approach would allow to understand the mechanisms and the consequences of gene expression evolution.

We have developed Bgee (a dataBase for Gene Expression Evolution) to address these questions. Bgee must answer the following requirements, to enable large scale gene expression pattern comparison:

- Precise description of the anatomy and developmental stages of each species, stored in a computer-understandable way.



- Integration of expression data in order to know in which anatomical features (spatial mapping) and which developmental stages (temporal mapping) genes are expressed.
- Comparison criteria between anatomies, developmental stages, and genes.

To unambiguously describe anatomy and development of a species in a computer-understandable way, ontologies are required: they describe a domain of knowledge, by using well-defined concepts and designing relationships amongst them. Several databases provide species-specific ontologies that describe anatomical features for a species, such as ZFIN [8] for the zebrafish. But as far as we know, no database provides relationships between these ontologies to allow comparisons.

The appropriate criterion to make comparisons in an evolutionary context is homology: we need to compare features that derive from the same ancestral element. We have thus designed homology relationships between anatomies of different species. This is a difficult task, and Bgee implements computational methods to achieve it (section 2). Then, we need homology relationships between genes. This point has already been abundantly treated in bioinformatics, and will not be discussed in detail in this paper. Finally, we need relationships between developmental stages. As these stages are artificial features that help to describe the continuous process of development, homology cannot be defined in a rigorous manner. We have rather designed a mapping of “equivalent” developmental stages between species (section 3).

To describe gene expression patterns, Bgee requires large amounts of data. To this end, heterogeneous data types are used (ESTs, microarrays, and soon *in situ* hybridizations). The common information to gather is whether an experiment has determined that a gene is expressed or not, and with which confidence. We have applied different statistical tests for each data type to obtain this information (section 4).

Thanks to the successful implementation of all these requirements (anatomical and developmental ontologies, comparison relationships between ontologies and genes, integration of heterogeneous expression data), Bgee allows the easy retrieval of gene expression data for different species, as well as the automated comparison of gene expression patterns.

## 1.2 Designing Homology Relationships between Anatomical Ontologies by an Ontology Alignment Approach

To study the evolution of gene expression patterns, comparisons have to be done between organs that evolved from a common ancestral structure. Thus designing relationships between anatomical ontologies consists in finding correspondences (homology relationships) between the concepts (organs) of these ontologies. This problem is a special case of “schema matching”, or “ontology alignment”.

Ontology alignment ([9] for a review) is the process of determining correspondences between ontology concepts. Usually, this technique is used to find the common concepts present in two ontologies. In the case of anatomical ontologies, the concepts to align are not strictly common, but rather, related: a homology relationship is not an equivalence relationship. For this reason, ontology alignment approaches developed for other applications cannot be applied as is: these methods would be misled by the existence of elements of same names and related to the same concept, but not homologous (eye of insects and of vertebrates for instance), or reciprocally, homologous elements with different names (pectoral fin and upper limb for instance). This is why we apply modified ontology alignment techniques in order to find putative homologies between two species anatomies. An expert has to manually validate the putative homologs. This method is implemented by Homolonto, a software that we have developed in Java. Homolonto will be presented in detail elsewhere; we present here the outline of its algorithm.

Our process is a supervised one: at each step, some homology relationships are proposed to the expert, who may validate them or not. Computations are made based on these decisions, and new propositions are made to the expert.

The algorithm starts with a list of pairs, which have identical names. This is based on the assumption that two structures that have the same name are likely homologous. For example, “optic cup” of the ZFIN ontology (zebrafish) and “optic cup” of the EHDA ontology (human) will be paired, but “optic cup” of ZFIN will not be initially paired with “optic nerve” of EHDA. The score of similarity between terms is up weighted by the proportion of common words, and down weighted by the frequency of these words (frequent words are less informative, e.g. “endoderm”). Moreover, scores are propagated between pairs which are neighbors in both ontologies. For example, the

score of the “optic cup” pair is added to the score of the “eye” pair, as “optic cup” is part of “eye”. In the same way, the score of the “eye” pair is added to the “optic cup” one.

Each pair is proposed to the expert, in descending order of scores. The expert may validate or invalidate the hypothesis of homology, or delay decision. The expert may choose to evaluate any number of pairs before triggering an iteration, in which computations are performed. Computations create or extend homology groups. The new homology information is propagated through the ontologies. The underlying idea is that if two concepts A and B are homologous, then one of the sub-concepts of A is probably homologous to one of the sub-concepts of B even if they have different names. Of note, validated homology contributes a significantly higher score than name similarity. Propagation is down weighted by the number of sub-concepts, to avoid generating many false positives (e.g. all the children of “whole body”).

Evaluation of pairs, ordered by total score (base score + propagated score), and iteration, are repeated until the expert decides to terminate, or no more pairs are proposed. Compared to manual alignment of the ontologies, Homolonto reduces time considerably, with high sensitivity. Thus aligning the zebrafish (ZFIN; 2087 terms) and Xenopus (Xenbase; 480 terms) ontologies took one month by hand, but 2 days using Homolonto. The first 213 pairs proposed to the expert were valid at 80%, and contained 91% of all true positives.

To design homology relationships between several species, we merge the homology groups obtained by pair-wise alignment.

Finally, Homolonto generates an OBO [10] file containing the homology relationships. Bgee then parses this file to integrate the homologies into the database.

### **1.3 Mapping of the Developmental Ontologies**

In relationship with the anatomical ontologies, Bgee uses for each species an ontology which describes its developmental stages, and links them using an *is\_a* relationship by key states (e.g. embryo, hatching, larval).

To compare expression patterns, the comparisons have to be done both between homologous organs (see section 2), and at an equivalent developmental stage. But it is not possible to “simply” identify stages between species for which the state of the development is identical: organs do not develop at the same speed and with the same sequence, development is heterochronous (e.g. [11]).

A solution could be to identify, for each organ involved in a homology relationship, the different key states of their formation, and to design, organ by organ, equivalence relationship between these states in different species. This solution is difficult to implement, as it would imply manual definition for each organ separately, without any guiding principle in the data (i.e. we cannot use shared names and ontology structures as for anatomical homology).

Although there is no direct equivalence between the stages of two species because of heterochrony, it is instead possible to identify key events of development, common to all bilaterian animals. We have developed a small ontology of these common “metastages”: embryo – including zygote, cleavage, blastula, gastrula, organogenesis –, post-embryonic development, adult. Then we have mapped the developmental stages of each species to these “metastages”. This approach results in a loss of accuracy regarding the developmental ontologies, but allows to compare gene expression patterns taking into account the time dimension.

## **1.4 Integrating Heterogeneous Data on Anatomical and Developmental Ontologies**

Integrating heterogeneous expression data is challenging, as it is difficult to compare the results of different types of techniques (e.g. ESTs, microarrays, *in situ* hybridizations) [12, 13], and even for a same type, to compare results between experiments (e.g. compare two microarray experiments made on different platforms). But as we want to be able to precisely describe expression patterns of genes, we need data as complete as possible. We also want to obtain data for all the species studied, and some techniques cannot be applied to all species, for instance *in situ* hybridizations on human. The information we want to collect is in which organs, and at which developmental stages, a gene is expressed. It means that for each experiment, we have to map the data to anatomical and developmental ontologies, and to apply statistical analyses, depending on the data type, to identify genes significantly expressed.

### ***Mapping Expression Data to Ontologies***

The main problem to map the data to ontologies is that annotations are often inconsistent between data sources: for instance, the description of the organs on which an experiment has been performed can be provided as free text, controlled vocabularies,

or ontologies. Therefore, we have manually annotated each experiment stored in Bgee to determine the unique identifiers (ID) in the anatomical ontologies of the organs studied, and the ID of the developmental stages.

The granularity of the data is also highly variable. For instance, experiments can be reported on the organ “brain” or on the organ “forebrain”, at the stage “embryo” or at the stage “free blastocyst”. This is why ontologies are essential both for anatomy and for development: just listing the developmental stages would not have been sufficient.

### ***Statistical Analyses***

Bgee currently uses EST data from Unigene [14] and Affymetrix data retrieved from ArrayExpress [15]. For each data type, Bgee applies statistical tests to identify genes that are significantly expressed, with two levels of confidence: low and high.

For experiments based on tag counting, such as EST, SAGE, or MPSS, a statistical test [16] shows that a gene is expressed with a 95% confidence if 7 tags are mapped to this gene (the number of tags is statistically different from 0). So for EST data, we have considered a gene as expressed with a high confidence if an experiment has found at least 7 EST related to this gene, and with a low confidence from 1 to 6 EST.

Affymetrix data are measurements of fluorescence intensity. Labelled cDNAs prepared from samples are hybridized with oligonucleotide probes. All probes mapping to the same transcript constitute a probeset. Identifying genes significantly expressed consists in finding genes for which the signal of the probeset is significantly different from the background signal. This method is implemented by the MAS5 software [17]; based on these statistical analyses, probesets are flagged as “present”, “marginal”, or “absent”. This allows us to classify genes expressed with a high confidence when their probeset is flagged as “present”, and with a low confidence when “marginal”. Although MAS5 classification is efficient [18], the estimation of the background signal can be biased depending on probe sequence affinity [19]. We are currently implementing another method of detection [19], which uses the gcRMA algorithm [20] to normalize the signal taking into account probe sequences, and uses a subset of weakly expressed probesets for estimating the background. A Wilcoxon test is then applied to compare the normalized signal of the probesets with the background signal. Genes will be considered expressed with a high confidence if the p-value is lower than 1%, and with a low confidence if the p-value is between 1 and 5 %.

Bgee will soon include *in situ* hybridization data. For data based on image analyses, statistical tests cannot be applied easily. Determining if a gene is expressed is usually done manually by an expert. A quality annotation can also be provided, summarizing the quality of the image, the hybridization, and the probes design. Such information is already present in several databases (e.g. ZFIN [8]), and Bgee will rely on them.

## 1.5 Database and Web-Interface of Bgee

The database of Bgee is developed with MySQL, and currently includes anatomical ontologies, developmental ontologies, and expression data for four species: human, mouse, zebrafish, and Xenopus:

- The anatomical ontologies come from eVoc [21] for human, Xspan [22] for human and mouse, MGD [23] for adult mouse, ZFIN [8] for zebrafish, and Xenbase [24] for Xenopus.
- EST data come from Unigene [14] and Affymetrix data from ArrayExpress [15]. *In situ* hybridization will be collected from specialized databases, as ZFIN or BGEM [25].
- Gene ontology [26] annotations and homology relationships between genes are recovered from Ensembl [27].
- Bgee currently includes a total of 104,881 genes. 51,277 have expression data, in 587 anatomical structures and 93 developmental stages.

The web interface of Bgee is developed in Java using the servlet container Tomcat, with a Model-View-Controller architecture. The user experience is improved by the use of AJAX technologies (Asynchronous Javascript And XML). The website of Bgee, available at <http://bgee.unil.ch/>, proposes several ways to easily retrieve or compare expression data:

- Querying the database: data can be queried for genes, gene families, anatomical structures, or developmental stages, based on their names, synonyms, abbreviations, identifiers, or descriptions.
- Browsing the ontologies: anatomical and developmental ontologies can be browsed as a tree structured view. Information about the genes expressed is displayed for each anatomical structure or developmental stages. The display of these expression data can be adjusted by selecting data type and data quality, or by entering a list of gene identifiers or of GO terms.

- Retrieving the expression pattern of a gene: the expression pattern of a gene is also displayed as a tree structured view of the organs where it is expressed, at the selected developmental stage. The data used to define the pattern can be modified by selecting the data type or data quality.
- Comparing the expression patterns of homologous genes: the expression patterns of a gene family can be compared choosing the species studied, and as for the ontology browsing, by selecting data type and quality, list of genes or of GO terms.

The homology relationships and developmental ontologies, both in OBO format, the Homolonto software and source code, and the Bgee database and source code, will soon be available on our website.

## 1.6 Conclusions

We have developed pipelines to integrate ontologies and expression data to Bgee, and automatically perform statistical analyses. We also have developed the Homolonto software to facilitate the design of homology relationships. We have paid great attention to make the Java code of Bgee easy to evolve, with a clean architecture and reusable components. We have thus implemented all the requirements to add more species and more data types into Bgee in the future. We plan to add in the short-term *in situ* hybridization data.

The multi-species computer coding and storage of expression patterns was an essential key to perform high throughput analyses. We will now be able to design analysis tools dedicated to the comparison of expression patterns, and to address open biological questions, such as the relationships between evolution of development and of gene expression, or the identification of candidate genes for diseases.

## 1.7 Acknowledgements

We thank Frederic Ricci for data annotation. Funding was provided by Etat de Vaud, the program Crescendo, the SIB, the Decryphon program.

## 1.8 References

1. Carroll, S.: *Endless Forms Most Beautiful: The New Science of Evo Devo and The Making of the Animal Kingdom*. W. W. Norton & Company, New York (2005)
2. Yanai, I., Graur, D., et al.: Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. *Omics* 8, 15-24 (2004)
3. Jordan, I.K., Marino-Ramirez, L., et al.: Evolutionary significance of gene expression divergence. *Gene* 345, 119-126 (2005)
4. Schlicht, M., Matysiak, B., et al.: Cross-species global and subset gene expression profiling identifies genes involved in prostate cancer response to selenium. *BMC Genomics* 5, 58 (2004)
5. Gu, Z., Nicolae, D., et al.: Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet* 18, 609-613 (2002)
6. Gu, X., Zhang, Z., et al.: Rapid evolution of expression and regulatory divergences after yeast gene duplication. *Proc Natl Acad Sci U S A* 102, 707-712 (2005)
7. He, X., Zhang, J.: Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. *Genetics* 169, 1157-1164 (2005)
8. Sprague, J., Clements, D., et al.: The Zebrafish Information Network (ZFIN): the zebrafish model organism database. *Nucleic Acids Res* 31, 241-243 (2003)
9. Shvaiko, P., Euzenat, J.: *Ontology Matching*. Springer Verlag, Berlin Heidelberg (2007)
10. Smith, B., Ashburner, M., et al.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 25, 1251-1255 (2007)
11. Jeffery, J.E., Bininda-Emonds, O.R., et al.: A new technique for identifying sequence heterochrony. *Syst Biol* 54, 230-240 (2005)
12. Lee, C.K., Sunkin, S.M., et al.: Quantitative methods for genome-scale analysis of in situ hybridization and correlation with microarray data. *Genome biology* 9, R23 (2008)
13. Kuo, W.P., Liu, F., et al.: A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies. *Nat Biotechnol* 24, 832-840 (2006)
14. Wheeler, D.L., Barrett, T., et al.: Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 36, D13-21 (2008)
15. Parkinson, H., Kapushesky, M., et al.: ArrayExpress--a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res* 35, D747-750 (2007)
16. Audic, S., Claverie, J.M.: The significance of digital gene expression profiles. *Genome Res* 7, 986-995 (1997)
17. Liu, W.M., Mei, R., et al.: Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics (Oxford, England)* 18, 1593-1599 (2002)
18. Choe, S.E., Boutros, M., et al.: Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome biology* 6, R16 (2005)
19. Schuster, E.F., Blanc, E., et al.: Correcting for sequence biases in present/absent calls. *Genome biology* 8, R125 (2007)



20. Wu, Z., Irizarry, R.A., et al.: A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association* 99, 909-917 (2004)
21. Kruger, A., Hofmann, O., et al.: Simplified ontologies allowing comparison of developmental mammalian gene expression. *Genome biology* 8, R229 (2007)
22. Aitken, S.: Formalizing concepts of species, sex and developmental stage in anatomical ontologies. *Bioinformatics (Oxford, England)* 21, 2773-2779 (2005)
23. Eppig, J.T., Blake, J.A., et al.: The mouse genome database (MGD): new features facilitating a model system. *Nucleic Acids Res* 35, D630-637 (2007)
24. Bowes, J.B., Snyder, K.A., et al.: Xenbase: a *Xenopus* biology and genomics resource. *Nucleic Acids Res* 36, D761-767 (2008)
25. Magdaleno, S., Jensen, P., et al.: BGEM: an in situ hybridization database of gene expression in the embryonic and adult mouse nervous system. *PLoS Biol* 4, e86 (2006)
26. Ashburner, M., Ball, C.A., et al.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* 25, 25-29 (2000)
27. Hubbard, T.J., Aken, B.L., et al.: Ensembl 2007. *Nucleic Acids Res* 35, D610-617 (2007)



## 2 An ontology to clarify homology-related concepts

---

Julien Roux, Marc Robinson-Rechavi

### **Abstract**

Although homology is a fundamental concept in biology, and is one of the shared channels of communication universal to all biology [1], it is difficult to find a consensus definition [2]. The interpretations of homology have changed as biology has progressed. New terms have been introduced into the literature, such as paramorphism [3], with mixed success. In addition, different research fields operate with different definitions of homology, for example the mechanistic usage of Evo-Devo [4] is not strictly historical, and would not be acceptable in cladistics. This makes a global understanding of homology complex, whereas the integration of evolutionary concepts into bioinformatics and genomics is increasingly important. We propose an ontology organizing homology and related concepts, which might provide a solution, and we hope it will also facilitate the integration and sharing of knowledge among the community.

This article was published in Trends in Genetics (2010) 26(3), 99-102

[doi:10.1016/j.tig.2009.12.012](https://doi.org/10.1016/j.tig.2009.12.012)

## **2.1 The problem: the concept of homology is divided by specialized usage**

The lack of a consensus definition of homology does not prevent us from perceiving and recognizing homologies in practice. Scientists have long been trying to understand the underlying cause of homology [1,2,5]. Several working definitions exist in specific fields of research. One example is the concept of homology based on common descent, applied at the molecular level. Many terms describing specific evolutionary histories of sequences, such as orthology or paralogy (Figure 1), are commonly used in genetics and molecular evolution.

But the abundance of terms has become another hurdle to a good understanding of homology related concepts for biologists [1]; most of them are redundant or very specialized. Importantly this terminological confusion can also hinder large-scale studies: in comparative and evolutionary biology, with the exponential increase of data available, the use of high-throughput computational tools is now generalized. There is a need for a bioinformatics framework to deal with the multiplicity of concepts related to homology.

## **2.2 Towards a solution: an ontology of homology related terms**

An ontology can provide such a framework. Ontologies are increasingly being used for data integration in biology [6] and can provide an efficient way to organize knowledge. Based on definitions from the literature, we have reviewed and organized terms related to the concept of homology into an ontology with an emphasis on the terms in modern use. This accounts for 65 terms plus 67 synonyms. The HOM ontology is presented according to Open Biological Ontologies Foundry principles [7] (<http://www.obofoundry.org/>), including a definition of each term and key references. The relationships between the terms are explicit, with some concepts as sub-classes of others (Figure 1). An overview of the type of information gathered is shown in Table 1; the full details can be obtained from the following website <http://www.ebi.ac.uk/ontology-lookup/browse.do?ontName=HOM>.

## ***Similarity as root***

An important choice when developing an ontology is the choice of the root (i.e. the most general term) because this defines the domain of application of the ontology. The root of the HOM ontology is ‘similarity’, or ‘sameness’. To quote Stevens: “without some similarity, we should not even dream of homology” [8]. We define it as a relation between biological objects that resemble or are related to each other sufficiently to warrant a comparison.

‘Homology’ is thus a sub-class of similarity. Another is ‘homoplasmy’ (or ‘analogy’, but the use of this term is ambiguous in the literature), describing similarity due to independent evolution. These two concepts are traditionally considered as disjoint or separate (although see Ref. [9]), and are defined as such in HOM.

Other sub-classes of similarity are independent of a homology hypothesis: ‘homocracy’ is the relation between two structures that share homologous patterning genes [10] and ‘functional equivalence’ is used to state that two structures share the same function.

## ***Working definitions of homology***

We propose a broad definition of homology, which encompasses the definitions proposed so far and can be seen as a common denominator or minimal approach: ‘similarity that results from common evolutionary origin’ [5].

Three different operational definitions, which are not disjoint [4], are gathered under this broad umbrella: (i) ‘Historical homology’ is the notion of similarity due to common descent [5]. (ii) ‘Biological homology’, fitting evo-devo usage, is process-oriented and more mechanistic, focusing on establishment and individualization of structures through common developmental processes [4]. It accommodates repeated parts of the same organism (‘iterative homology’) and sexually differentiated parts of individuals (e.g. testis and ovaries). (iii) ‘Structural homology’ refers to the traditional criteria of homology focused on similarity with regard to selected structural parameters (sometimes called ‘idealistic homology’ [1,4]).

## ***Multiple inheritance***

An ontology can represent complex concepts by encoding multiple inheritance: a term can be a sub-class of more than one other term. Examples where homology statements do not translate smoothly between multiple levels of organization (e.g. anatomical

structures and genes) are easily represented. For example, ‘deep homology’ is a subclass both of homoplasy and of homocracy, because it involves anatomical structures that result from independent evolution and yet share the expression of homologous patterning genes [11].

### ***Availability***

The HOM ontology is available at [www.obofoundry.org](http://www.obofoundry.org). Interactive views are available at the Bioportal (<http://bioportal.bioontology.org/ontologies/40983/>, see Figure 1) or the Ontology Lookup Service at EBI (<http://www.ebi.ac.uk/ontology-lookup/browse.do?ontName=HOM>).

## **2.3 Concluding remarks**

Discussions related to the concept of homology have led to repeated confusion. Like discussions on the terms ‘species’ or ‘gene’, it is not clear whether a better understanding will simply emerge from future advances in biology. Indeed, what makes the concept intrinsically difficult to outline is probably the complexity of living organisms and their evolution. As West-Eberhard puts it: “evolution makes a mess of homology” [12].

In this context, we feel that the most helpful solution is to order and clarify existing concepts. This should provide an evolvable tool for computational studies, and a framework for future conceptual advances (i.e. proposals for new terms should be set in relation to existing concepts).

## **2.4 Acknowledgments**

We acknowledge funding from Etat de Vaud and Swiss National Science Foundation grant 116798. We thank Frederic Bastian and other members of the lab for discussions; the members of the OBO foundry mailing list for their suggestions; and the National Center for Biomedical Ontology for authorization of using their visualization tool in Figure 1. We apologize to the many colleagues whose work could not be cited because of space limitations; all citations relevant to specific terms are included in the HOM ontology.

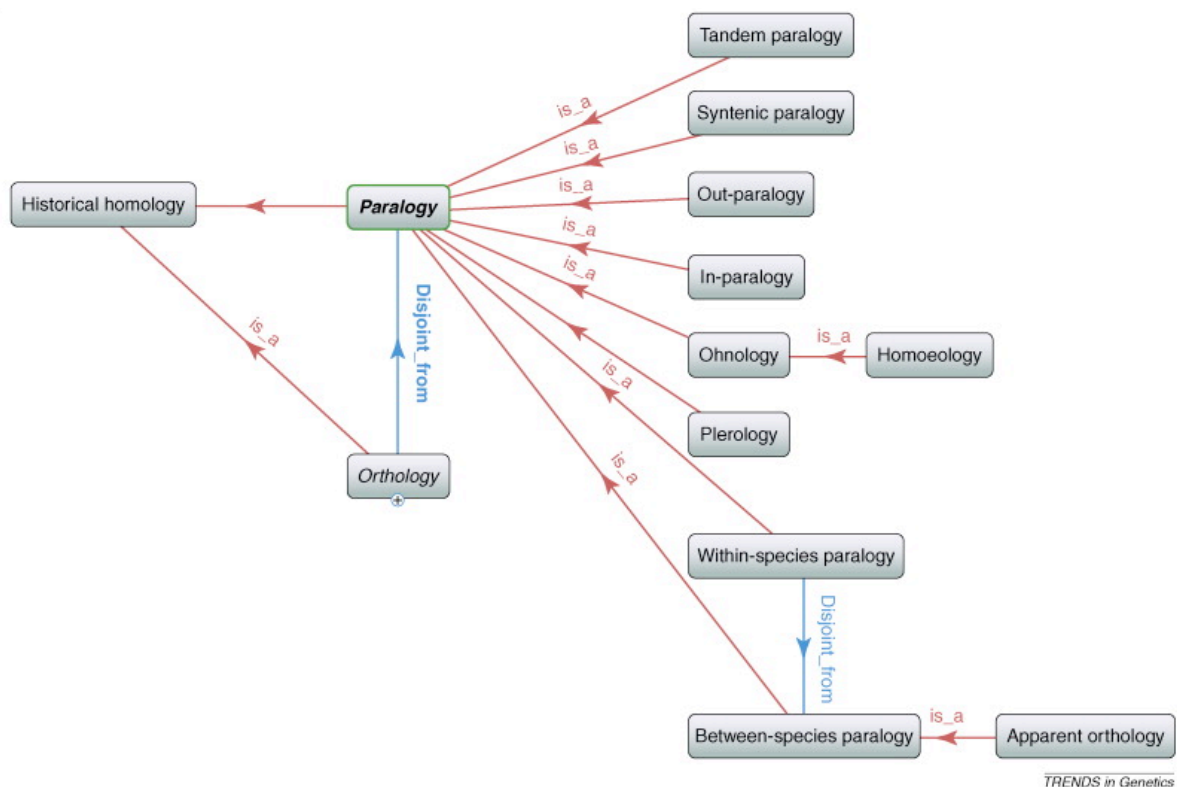


Figure 1. A partial view of the ontology of homology and related concepts (HOM). The concepts related to the concept of 'paralogy' are displayed. Boxes represent terms, arrows represent relations between the terms. The relation 'is\_a' denotes that one term is a sub-class of another. Courtesy of the National Center for Biomedical Ontology. Copyright © 2005–2009, Stanford University. <http://bioportal.bioontology.org> and <http://keg.cs.uvic.ca/ncbo/flexviz/FlexoViz.html>

Table 1. Example of data represented in HOM ontology for 'paralogy' and 'latent homology'

	Example 1	Example 2
<b>Id</b>	HOM:0000011	HOM:0000057
<b>Name</b>	Paralogy	Latent homology
<b>Definition</b>	Historical homology that involves genes that diverged after a duplication event.	Parallelism that involves morphologically very similar structures, occurring only within some members of a taxon and absent in the common ancestor (which possessed the developmental basis to develop this character).
<b>is_a</b>	HOM:0000007 historical homology	HOM:0000005 parallelism HOM:0000058 syngeny
<b>References</b>	<p>Fitch WM (2000) Homology: a personal view on some of the problems. Trends in Genetics 16:227-231.  <a href="https://doi.org/10.1016/S0168-9525(00)02005-9">doi:10.1016/S0168-9525(00)02005-9</a></p> <p>Fitch WM (1970) Distinguishing homologous from analogous proteins. Syst. zool. 19(2): 99-113.  <a href="https://pubmed.ncbi.nlm.nih.gov/5449325/">PMID:5449325</a></p> <p>Koonin EV (2005) Orthologs, paralogs, and evolutionary genomics. Annual Review of Genetics, 39: 309-338.  <a href="https://doi.org/10.1146/annurev.genet.39.073003.114725">doi:10.1146/annurev.genet.39.073003.114725</a></p>	<p>Rutishauser R and Moline P (2005) Evo-devo and the search for homology ("sameness") in biological systems. Theory in Biosciences 124:213-241.  <a href="https://doi.org/10.1007/BF02814485">doi:10.1007/BF02814485</a></p> <p>Hall BK (2007) Homoplasy and homology: Dichotomy or continuum? Journal of Human Evolution. 52:5, 473-479.  <a href="https://doi.org/10.1016/j.jhevol.2006.11.010">doi:10.1016/j.jhevol.2006.11.010</a></p> <p>Sanetra M et al. (2005) Conservation and co-option in developmental programmes: the importance of homology relationships. Frontiers in Zoology 2:15.  <a href="https://doi.org/10.1186/1742-9994-2-15">doi:10.1186/1742-9994-2-15</a></p> <p>de Beer G (1971). Homology, an unsolved problem. London, Oxford University Press.  <a href="https://www.isbn-international.org/en/ISBN:0199141118">ISBN:0199141118</a></p>
<b>Cross-references</b>	SO:0000854 paralogous_region SO:0000859 paralogous SO:paralogous_to	
<b>Comment</b>		Used for structures in closely related taxa
<b>Synonyms</b>		Apomorphic tendency (exact) Cryptic homology (exact) Homoiology (related) Homoplastic tendency (related) Re-awakening (related) Underlying synapomorphy (exact)



## 2.5 References

1. Kleisner K (2007) The Formation of the Theory of Homology in Biological Sciences. *Acta Biotheoretica* 55: 317-340.
2. Brigandt I (2003) Homology in comparative, molecular, and evolutionary developmental biology: The radiation of a concept. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 299B: 9-17.
3. Minelli A (2000) Limbs and tail as evolutionarily diverging duplicates of the main body axis. *Evol Dev* 2: 157-165.
4. Wagner GP (1989) The Biological Homology Concept. *Annual Review of Ecology and Systematics* 20: 51-69.
5. Hall B, editor (1994) *Homology: the hierarchical basis of comparative biology*: Academic Press. 483 p.
6. Bard JB, Rhee SY (2004) Ontologies in biology: design, applications and future challenges. *Nat Rev Genet* 5: 213-222.
7. Smith B, Ashburner M, Rosse C, Bard J, Bug W, et al. (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotech* 25: 1251-1255.
8. Stevens P (1984) Homology and phylogeny: morphology and systematics. *Systematic botany* 9: 395-409.
9. Hall BK (2007) Homoplasy and homology: Dichotomy or continuum? *Journal of Human Evolution* 52: 473-479.
10. Nielsen C, Martinez P (2003) Patterns of gene expression: homology or homocracy? *Development Genes and Evolution* 213: 149-154.
11. Shubin N, Tabin C, Carroll S (1997) Fossils, genes and the evolution of animal limbs. *Nature* 388: 639-648.
12. West-Eberhard M (2003) *Developmental plasticity and evolution*: Oxford University Press, USA. 816 p.



# 3 Developmental Constraints on Vertebrate Genome Evolution

---

Julien Roux, Marc Robinson-Rechavi

## Abstract

Constraints in embryonic development are thought to bias the direction of evolution by making some changes less likely, and others more likely, depending on their consequences on ontogeny. Here we characterize the constraints acting on genome evolution in vertebrates. We use gene expression data from two vertebrates: zebrafish, using a microarray experiment spanning 14 stages of development, and mouse, using EST counts for 26 stages of development. We show that, in both species, genes expressed early in development (i) have a more dramatic effect of knock-out or mutation, and (ii) are more likely to revert to single copy after whole genome duplication, relative to genes expressed late. This supports high constraints on early stages of vertebrate development, making them less open to innovations (gene gain or gene loss). Results are robust to different sources of data: gene expression from microarrays, ESTs or *in situ* hybridizations; mutants from directed KO, transgenic insertions, point mutations, or morpholinos. We determine the pattern of these constraints, which differs from the model used to describe vertebrate morphological conservation (“hourglass” model). While morphological constraints reach a maximum at mid development (the “phylotypic” stage), genomic constraints appear to decrease in a monotonous manner over developmental time.

This article was published in PLoS Genetics (2008) 4(12): e1000311.

[doi:10.1371/journal.pgen.1000311](https://doi.org/10.1371/journal.pgen.1000311)

### 3.1 Introduction

To what extent do the processes of embryonic development constrain genome evolution? Correlations between developmental timing and morphological divergence have long been observed, but the mechanisms and molecular basis of such patterns are poorly understood. The most commonly used measure of selective pressure on the genome, the ratio of non-synonymous to synonymous substitutions ( $d_N/d_S$ ) in protein coding genes, has been of limited help in this case. Stronger constraints have been found on genes expressed in late embryonic stages in *Drosophila* [1], but most other studies have failed to report robust evidence for a lower  $d_N/d_S$  ratio in genes expressed at constrained developmental stages [2-5]. A different approach has been to characterize which genes are duplicated, and which are not: studies of *C. elegans* [2] and *Drosophila* [6] have found less duplication of genes expressed in early development. These results show that it is possible to identify developmental constraints at the genomic level. They have a few limitations though. One is that the data available has limited the characterization of developmental time to broad categories such as “early” and “late”. A second is the difficulty of relating results from two derived invertebrate species, to morphological evolution models in vertebrates [7].

Indeed it is in vertebrates that the fundamental models of developmental constraint on evolution have been established, starting in the nineteenth century with the “laws” of von Baer [8], claiming a progressive divergence of morphological similarities between vertebrate embryos, with the formation of more general characters before species-specific characters. Integration of these observations within evolutionary biology has not always been straight-forward [9-11]. More recently, an “hourglass” model was proposed to describe morphological evolution across development [12,13]: in the earliest stages of development (cleavage, blastula) there is in fact a great variety of forms in vertebrate embryos. Later in development, a “phylotypic” or conserved stage is observed, where many morphological characteristics are shared among vertebrates. This stage is usually presumed to be around the pharyngula stage. After this bottleneck, a “von Baer-like” progressive divergence is again observed. The conserved phylotypic stage has been explained by assuming higher developmental constraints [13-15]. The limits on morphological evolution would be placed by the structure of animal

development, making some changes unlikely or impossible. How such limitations are encoded in the genome, or impact its evolution, is still an open question.

In this work, we investigate the existence and timing of constraints on genes expressed in vertebrate development. We use representatives of the two main lineages of vertebrates, a teleost fish and a tetrapode, and we explore the impact of experimental gene loss, and of gain of gene copies in evolution. We find that timing of development has a strong impact in both cases, but that the pattern of constraints on genome evolution does not follow the morphological hourglass model. High constraints are present in early stages of development and relax progressively over time.

## 3.2 Results

### *Constraints on gene loss-of-function in zebrafish*

First, we used the phenotypes of gene loss-of-function as an indicator of selective pressure on genes. We extracted genes essential for the viability of the zebrafish, giving a lethal phenotype when non functional [16]. We expect that the loss of a gene should be more deleterious if this gene is expressed at a developmental stage with strong constraints. Thus we estimated whether genes were expressed or not at each stage, and computed the ratio of expressed essential genes to expressed reference genes (no reported loss of function phenotype). We then plotted the variation across development of this ratio. We used two different types of data to evaluate the presence of gene expression: (i) expression patterns from *in situ* hybridizations (Figure 1A), and (ii) “present” or “absent” calls from an Affymetrix microarray experiment (Figure 1B). Results are consistent for both data types: the proportion of essential genes is higher among genes expressed in early development, with a significant negative correlation. For the *in situ* hybridizations (Figure 1A), a linear regression is significant, but a parabola is not. The parabola has been suggested as the quantitative expectation of an hourglass-like model [3,17]. These results indicate a continuous trend over developmental time, with stronger constraints on early development.

Considering gene expression either “present” or “absent” allows straightforward statistical analysis, but it is a strong approximation of the continuous nature of gene

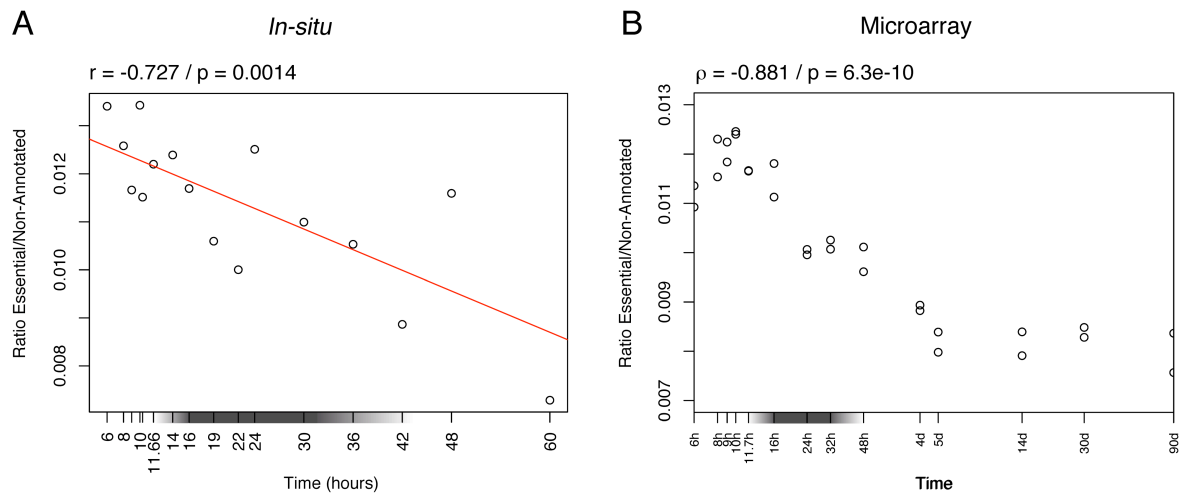


Figure 1: Variation across zebrafish development of the expression of essential genes compared to non-annotated genes. At each time point, the ratio of the number of essential genes expressed on the number of non-annotated genes expressed is plotted. A gray box on the x-axis indicates the phylotypic period. (A) Gene expression as reported using in situ hybridization data. The x-axis is proportional to time. A weighted linear regression was fitted to the data and the regression line plotted. (B) Gene expression as reported by “present” calls from Affymetrix array data. The x-axis is in logarithmic scale. A Spearman correlation was computed (coefficient  $\rho$ ).

[doi:10.1371/journal.pgen.1000311.g001](https://doi.org/10.1371/journal.pgen.1000311.g001)

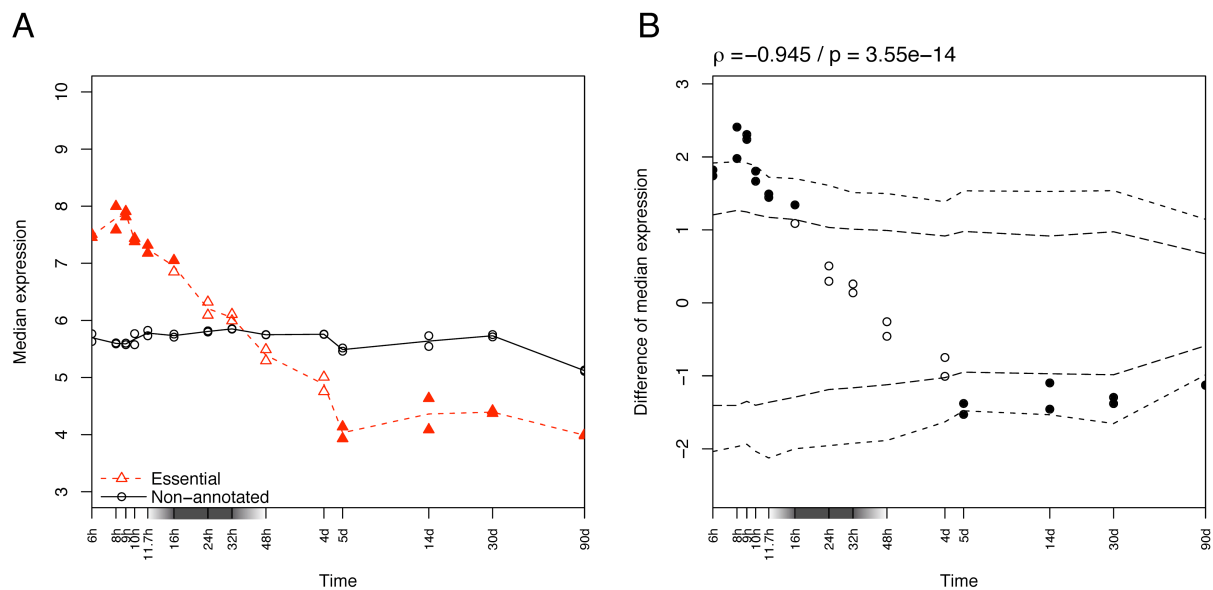


Figure 2: Expression in zebrafish development of essential genes. (A) Median expression profiles of zebrafish essential genes, in red dashed line and triangles, compared to non-annotated genes in black solid line and circles. (B) Significance of the expression difference between the two groups of genes. 1% and 1% confidence intervals are drawn in dashed lines. Significant points (outside the 1% confidence interval) are filled on both plots. A Spearman correlation was computed (coefficient  $\rho$ ) to test the trend over time. The x-axis is in logarithmic scale. A gray box on the x-axis indicates the phylotypic period.

[doi:10.1371/journal.pgen.1000311.g002](https://doi.org/10.1371/journal.pgen.1000311.g002)

expression. To take advantage of the quantitative signal from the microarray data, we contrasted the median expression level of all the essential genes to that of all of the reference genes (Figure 2A). We used the median because it is less sensitive to extreme values [18]; results were consistent using the mean (not shown). To estimate the significance of the difference between the two curves, we performed a randomization test (see Methods), which provides 1% and 1‰ confidence intervals (Figure 2B). The expectation is now that the essential genes should be enriched in genes highly expressed at the stages with strong constraints. And consistently with the previous observations, essential genes are significantly more expressed in early stages (until 11.7 hours), and less expressed in late stages of development (from 5 days to 14 days). No specific trend is visible around the phylotypic stage. Similar results are obtained for genes which give an “abnormal” phenotype after loss of function (Text S1 and Figure S4).

To complement this approach, we defined groups of genes according to their expression pattern during development (see Methods). This clustering of zebrafish genes provided us notably with a cluster of 2446 genes with high expression in early development, decreasing over time (Figure 3, cluster 3), and an opposite cluster of 1123 genes lowly expressed in early development, increasing over time (Figure 3, cluster 4). As expected, genes whose expression is highest in early development are more frequently essential (1.1% vs. 0.6%), and induce more frequently abnormal phenotypes when non functional (6.1% vs. 2.9%).

### ***Constraints on gene loss-of-function in mouse***

We performed a similar analysis in mouse, with some differences of methodology due to the data available. For expression, we used of a large amount of EST (Expressed Sequence Tags) data from libraries spanning development, from which we deduced presence or absence of expression (see Methods). Only phenotypes obtained by the targeted knock-out technique were used. As knock-out experiments with no observable phenotype are reported in mouse, we can use these as a reference set, instead of non annotated genes as in zebrafish. The ratio of expressed essential genes to expressed reference genes is significantly negatively correlated with developmental time (Figure 4A), as in zebrafish (Figure 1).

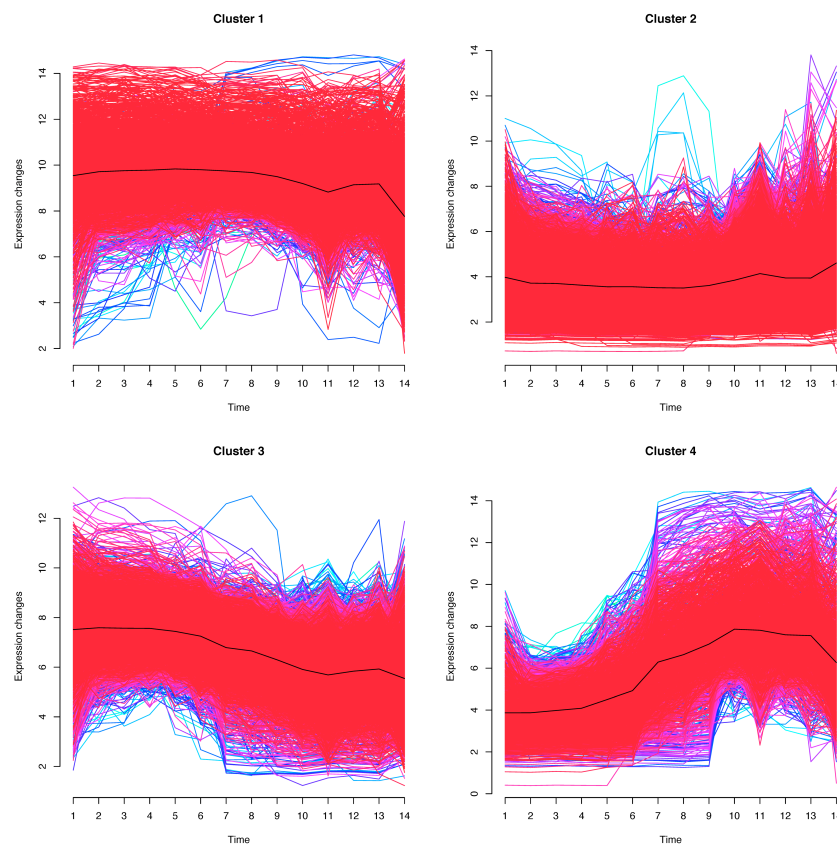


Figure 3: Expression of four groups of genes, clustered according to their expression in zebrafish development. [doi:10.1371/journal.pgen.1000311.g003](https://doi.org/10.1371/journal.pgen.1000311.g003)

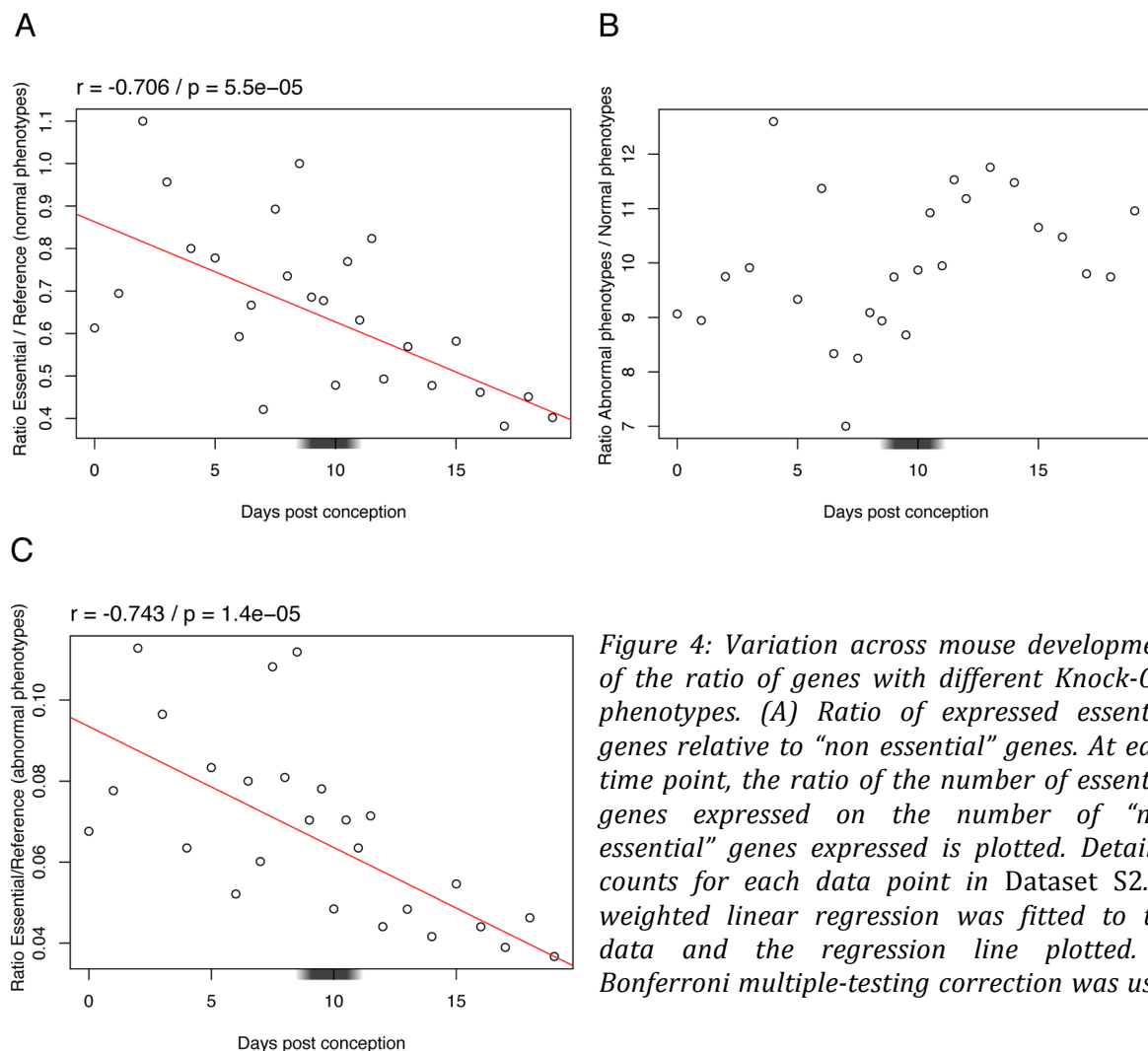


Figure 4: Variation across mouse development of the ratio of genes with different Knock-Out phenotypes. (A) Ratio of expressed essential genes relative to “non essential” genes. At each time point, the ratio of the number of essential genes expressed on the number of “non essential” genes expressed is plotted. Detailed counts for each data point in Dataset S2. A weighted linear regression was fitted to the data and the regression line plotted. A Bonferroni multiple-testing correction was used



to adjust the significance threshold ( $\alpha=0.05/6=0.0083$ ). A gray box on the x-axis indicates the phylotypic period. (B) Ratio of expressed genes inducing abnormal phenotypes when non functional compared to non essential genes. The linear regression is not significant after multiple testing correction ( $r = -0.477$ ;  $p = 0.014$ ). (C) Ratio of expressed essential genes compared to genes inducing abnormal phenotypes when non functional. Legend as in Figure 4A.

[doi:10.1371/journal.pgen.1000311.g004](https://doi.org/10.1371/journal.pgen.1000311.g004)

Repeating the same approach with genes inducing a phenotype reported as “abnormal” when they are not functional, no significant trend is detected compared to genes inducing no phenotype, after multiple testing correction (Figure 4B). Moreover, these genes can be used as a reference for essential genes (Figure 4C), with results very similar to the use of genes inducing no phenotype after loss of function (Figure 4A). Thus in mouse, genes inducing abnormal phenotypes when non-functional have a behavior more similar to the reference set of “non essential” genes.

### ***Constraints on gene duplication***

The fish specific whole genome duplication [19] provides us with a natural experiment on constraints on gene doubling: after this event approximately 85% of duplicated genes lost one copy, and the subset which retained both copies is known to be biased relative to function and selective pressure [20]. Thus we tested if duplicate gene expression pattern in zebrafish development was biased compared to singletons. We plotted the median expression profiles of duplicates originating from the fish specific whole genome duplication, and of singletons, genes whose duplicate copy has been lost after the genome duplication (Figure 5). Duplicates are less expressed in early stages of development. The difference of median expression decreases progressively, similar to the observations for essential or abnormal phenotype genes. Larval time points show a maximum expression of duplicates relative to singletons.

Two scenarios can explain this result. First, retention of two copies may be more likely after the whole genome duplication for genes less expressed in early development. Second, the retention of genes may be unbiased relative to development, but duplicate genes may evolve secondarily lower expression in early development. To get a proxy of the ancestral state before whole genome duplication, we used again mouse data, which has diverged from zebrafish before the fish specific duplication. We compared mouse orthologs of zebrafish duplicates to mouse orthologs of zebrafish singletons, regarding

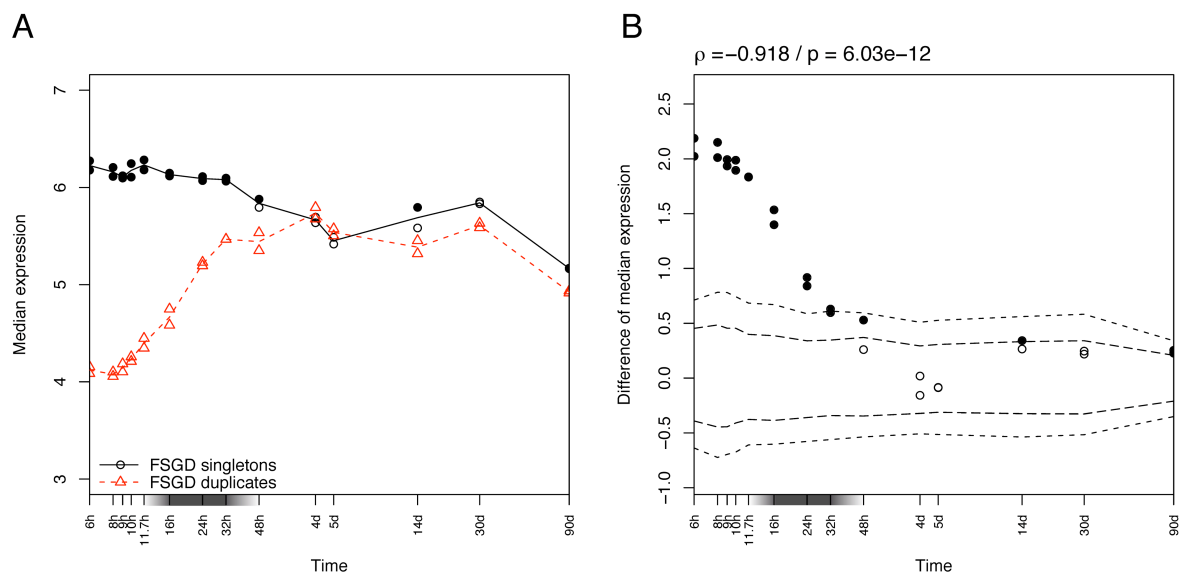


Figure 5: Expression in zebrafish development of genes according to retention after the fish specific whole genome duplication. Median expression profiles of zebrafish duplicates from the fish specific whole genome duplication in red dashed line and triangles, and of singletons in black solid line and circles. Legend as in Figure 2.

[doi:10.1371/journal.pgen.1000311.g005](https://doi.org/10.1371/journal.pgen.1000311.g005)

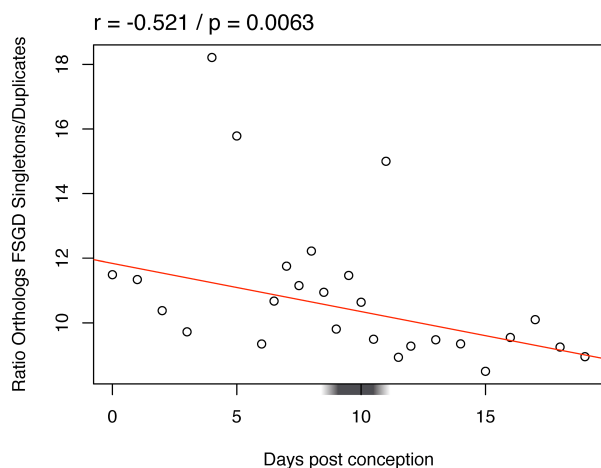
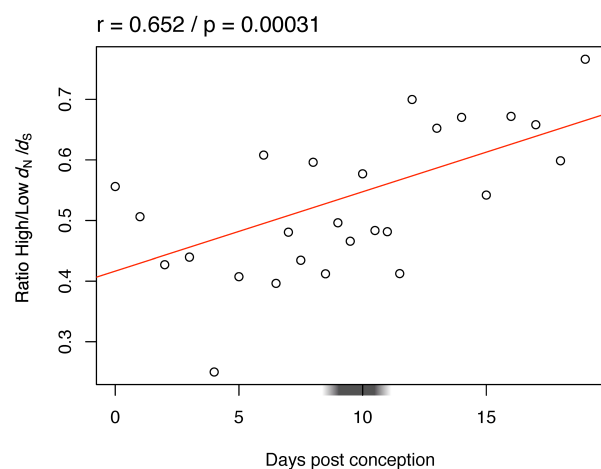


Figure 6: Variation across mouse development of the ratio of expressed orthologs of zebrafish singletons after the fish specific genome duplication (FSGD) relative to orthologs of zebrafish duplicates. Legend as in Figure 4.

[doi:10.1371/journal.pgen.1000311.g006](https://doi.org/10.1371/journal.pgen.1000311.g006)

Figure 7: Variation across mouse development of the expression of rapidly evolving genes (25% highest  $d_N/d_S$ ) compared to slowly evolving genes (25% lowest  $d_N/d_S$ ). Only singletons for 2R were considered. Legend as in Figure 4.

[doi:10.1371/journal.pgen.1000311.g007](https://doi.org/10.1371/journal.pgen.1000311.g007)



their expression in development (Figure 6). Mouse orthologs of duplicates are significantly less expressed in early development compared to orthologs of singletons. This result in mouse is consistent with the observations in zebrafish, and the most parsimonious explanation is that expression was similar in the ancestor of the two lineages. Therefore we can accept the first hypothesis: after the fish specific whole genome duplication, there was preferential retention of duplicates less expressed in early development.

To check if this phenomenon is particular to the fish specific genome duplication, we repeated this analysis with the two ancient rounds of genome duplication (“2R”), which occurred in the ancestor of vertebrates [21]. It is difficult to distinguish between the two whole genome duplications since no model species diverged from the vertebrate lineage between them. Therefore we looked at the median expression profiles of genes with any duplication at the origin of vertebrates, compared to singletons, whose duplicates were lost after both whole genome duplications. For zebrafish, we restricted this analysis to genes which are singletons regarding the fish specific whole genome duplication. Similarly to fish specific duplicates, duplicates from 2R are significantly less expressed than singletons in the early development of zebrafish (Figure S1) and mouse (Figure S2). Thus mechanisms of retention after whole genome duplication seem to be conserved during vertebrate evolution (see also Text S1).

### ***Constraints on gene sequence***

To check if sequences of genes expressed at different stages in development are experiencing different selective pressure, we used the non synonymous to synonymous substitution ratios ( $d_N/d_S$ ). In zebrafish, we used an approach similar to Davis et al. [1]: at each stage we performed the correlation between  $d_N/d_S$  and gene expression from microarray data (Figure S3). It has been shown that genes retained in duplicate tend to evolve slowly [20,22]. To control for that factor, we kept only strict singletons in the analysis (genes whose duplicate was lost after 2R and fish-specific genome duplications). At all stages the correlation is negative, confirming that genes with higher expression levels are under stronger purifying selection [23,24]. We note that correlation at the “adult” stage (90 days) is weaker (Figure S3): the link between

expression and selective constraints on sequences appears stronger in development than in adult. But there is not a significant trend over time (Spearman  $\rho = 0.08$ ;  $p = 0.68$ ). In mouse, we considered only singletons after 2R genome duplication, and we compared the slowest evolving genes (25% lower  $d_N/d_S$ ) with the fastest evolving genes (25% higher  $d_N/d_S$ ). There is a significant correlation with time of expression (Figure 7). Genes with strong sequence constraints (low  $d_N/d_S$ ) tend to be expressed early in development.

### ***Gene ontology characterization***

What is the function of the genes whose evolution is constrained by expression in early development? We analyzed enrichment or depletion in Gene Ontology [25] categories for the clusters based on gene expression (Figure 3). Using the Molecular Function ontology, genes whose expression is highest in early development are significantly enriched in fundamental processes of the cell, such as RNA processing, transcription, and DNA replication (Table S1). This is very similar to the categories observed to be enriched in house keeping genes [26]. It is also consistent with the categories depleted in fish specific duplicates [20]. Conversely, genes highly expressed in early development are depleted in receptor or channel activity, while these activities are enriched in genes highly expressed in late development. Fewer terms are significant for the Biological Process ontology, and results are essentially consistent with the Molecular Function. Overall, the genes expressed in early development, which appear constrained against gene duplication or loss of function, seem to be house keeping genes involved in basic cellular processes.

## **3.3 Discussion**

Recent discussion of the evolution of ontogeny [27] has allowed the clarification of several important points. The first is that models must be explicitly defined, to allow testing. Poe and Wake [17] distinguish three models for the evolution of ontogeny: the early conservation model *à la* von Baer [8]; the hourglass model, characterized by a conserved phylotypic stage [12,28]; and the adaptive penetrance model (an inverted hourglass). The second point is that quantitative testing is important to distinguish between these models. At the morphological level, several studies have used heterochrony data from vertebrates to quantify the amount of change at each stage of

development [17,29]. Surprisingly, this led to rejection of both the early conservation and the hourglass models, although which model is favoured remains disputed [27]. The third point that should be clarified is the distinction between constraints at the level of patterns, and constraints at the level of processes [29]. The studies of heterochrony in vertebrates are typically concerned with the pattern.

In this framework, our results clearly provide a quantitative test which supports the early conservation model. By studying not morphological structures but features of the genome and its expression, this test concerns the level of processes, not patterns. Thus an important point to be made is that our results should be taken neither in contradiction nor in support of any specific model at the level of patterns, given our still limited knowledge of causal relationships between process and patterns in ontogeny [30]. On the other hand, our results do appear to be in contradiction with previous reports of a maximum of constraints on processes around the phylotypic stage of vertebrates [3,4,31].

We use two simple measures of constraint on the expression of a gene at a developmental stage: if expression of one copy is needed, then (i) removing it may be deleterious, and (ii) increasing the number of copies may also be deleterious. This view is consistent with a recent study in yeast which suggests that constraints influencing the ability to lose certain genes or to maintain them in duplicate may be similar [32]. We expect gain or loss of genes highly expressed at more constrained developmental stages to be counter-selected. And indeed, we find a clear and significant trend: early development is strongly constrained, then constraints diminish during development in a continuous manner. Genes highly expressed in early development are more frequently essential, and less frequently preserved in double copy after genome duplication. Thus early development is less robust against gene loss and against gene doubling. Trends are conserved between mouse and zebrafish, representatives of the two main lineages of bony vertebrates, and between 2R and fish specific genome duplications. An indication of how strong these constraints are is our capacity to predict which genes were kept in duplicate in zebrafish based on expression pattern in mouse. Despite more than 400 MY of independent evolution, and the use of relatively noisy data (mix of EST libraries), more than a quarter of the variance in gene retention is explained (Figure 6;  $r^2 = 0.27$ ).

There is also some signal for early conservation at the level of coding sequences, at least in mouse (Figure 7). What we do not see is any genomic evidence for specific constraints at a phylotypic stage. Both in zebrafish and in mouse, the pharyngula stage appears to be part of the general trend from stronger genomic constraints in early development, towards weaker genomic constraints at later stages. We believe that our data are sufficiently detailed, and exhibit sufficiently strong signal, that a maximum of genomic constraints at the phylotypic stage would be visible. So where does the contradiction with previous studies come from?

An early quantitative study [31] found that when screens were done in rodents for the induction of teratogenesis, most abnormalities were obtained by applying teratogens during the phylotypic stage. This was interpreted [31] as supporting strong constraints at the phylotypic stage, due to inductive interactions. But these screens aimed not to test developmental robustness, but to obtain abnormal embryos for experimental work. As remarked by Bininda-Emonds et al. [29], Galis and Metz [31] define the phylotypic stage broadly as including most organogenesis. If application of teratogens in early development resulted in lethality before organogenesis, it would not be of interest to the researchers performing the screens. Thus it seems that what Galis and Metz [31] measured was the potential for a stage to produce morphological abnormalities, not the overall constraints on ontogeny at each stage. There seems to be little reason to suppose that such data provide “an accurate model of natural selection” [33], unlike e.g. the retention of duplicate genes over long evolutionary periods.

It is worth noting that we observe a “peak” of constraints shortly after pharyngula (Figure 4B) for the expression profile of mouse genes which give an “abnormal” phenotype when knocked-out. The behavior of these genes is surprising, because in zebrafish the trend for such genes was similar to that for essential genes. We suspect that the definition of abnormal phenotypes differs between databases and between investigators working in different species. Less severe phenotypes may be reported as “abnormal” in mouse, relative to zebrafish. Of note, data in ZFIN [16] come mainly from the reviewed literature, where minor abnormalities of phenotype are rarely reported, whereas data in the MGD [34] come also from genome wide mutagenesis, and thus include such minor abnormalities. Minor abnormalities in mouse phenotype may also be

easier to detect because of the gross similarity with human in anatomy and physiology. In any case, these are the data in our study which most closely approximate the teratogenesis study, and the only data that do not support the early conservation model. Although this trend is statistically not significant, it is consistent with the observations of Galis and Metz [31]. This deserves to be further examined in future studies.

Two other studies which quantified a maximum of constraints at the phylotypic stage did use evolutionary measures of constraint. These studies [3,4] estimated constraints on the evolution of coding sequences, in relation to the timing of expression in mouse development from EST data. Despite similar experimental designs and data, we reached differing conclusions. First, we note that we did check for sequence conservation ( $d_N/d_S$ ) trends over development. In zebrafish, we found no robust pattern (Figure S3), while in mouse we found support for the early conservation model (Figure 7). Second, in our analyses we found that small samples of ESTs could introduce important variability, which is why we used weighted regressions for all computations based on these data. For example, we see a very high ratio of mouse orthologs of zebrafish singletons to duplicates for Theiler stage 5 (day 4) (Figure 6); but this is obtained based on only 628 genes with at least one EST at that stage (median over all stages: 3767). The weighted regression insures that such a point has a weak incidence on the statistical significance. Similar issues are visible in the data of Irie et al. [4], but are not addressed in their analysis. Indeed, the extreme points they use to support constraints at pharyngula are based on some of the smallest samples of their dataset. Finally, it should be noted that another study in mouse found an opposite pattern (relaxation of constraints near the phylotypic stage) using an alternative measure of constraints on sequences, the ratio of radical to conservative amino acid changes,  $K_R/K_C$  [5]. In our opinion, these contradictory and weakly supported results are consistent with the idea that overall, coding sequence change seems to have a rather modest contribution to the evolution of development. This is consistent with a stronger contribution of regulation of expression [35,36].

Our results were obtained on data which either reflect the action of natural selection (duplicate gene retention), or are directly relevant to fitness (loss-of-function lethality), and provide unambiguous trends with strong statistical support. Moreover, the

consistent patterns in zebrafish *in situ* hybridization and microarray data, and mouse EST data, show robustness to potential experimental biases or sampling errors. The early conservation model for genomic processes is reinforced by the enrichment of early expressed genes in fundamental cellular processes (Figure 3; Table S1). This is the opposite of duplicated genes, which may be more involved in innovation, and have been reported to be enriched in developmental or behavioural processes [20,21]. Our results are consistent with the observation that basic cores of gene regulatory networks (GRNs) are highly constrained in early stages of animal development [37,38], although we add the notion of a progressive decrease in constraints. This indicates that some relations between the timing of cell-fate decisions in development and rates of genome evolution may be widely shared among animals [7,39]. Indeed, many studies underline gastrulation as a crucial step in development [40,41]. Accordingly this period is shown here to be subject to highest constraints, consistent with the famous Lewis Wolpert quote: "It is not birth, marriage, or death, but gastrulation, which is truly the most important time in your life" [42].

### 3.4 Materials and Methods

#### *Microarray data*

Microarray data of zebrafish (*Danio rerio*) development were downloaded from ArrayExpress (E-TABM-33) [43]. This experiment uses an Affymetrix GeneChip Zebrafish Genome Array (A-AFFY-38). 15 stages were sampled, spanning from fertilization to adult stages (15 minutes, 6, 8, 9, 10, 11.7, 16, 24, 30 hours, 2, 4, 5, 14, 30, 90 days, covering zygote, segmentation, gastrula, pharyngula, hatching, larval, juvenile, adult). Two replicates were made per time point; we use both of them for computations, and the 2 values are plotted to give an order of the variability between replicates.

Raw CEL files were renormalized using the package gcRMA [44] of Bioconductor version 2.2 [45]. We used the "affinities" model of gcRMA, which uses mismatch probes as negative control probes to estimate the non-specific binding of probe sequences. The normalized values of expression are in log2 scale, which attenuates the effect of outliers. Mapping of *D. rerio* genes on Affymetrix probesets was made using Ensembl [46] annotation for zebrafish genome version Zv7 (unpublished).

We did not consider the first time point of the data (15 minutes, fertilization). Its behaviour was peculiar in many cases. We explain this by the presence of maternal



transcripts in the embryo [47]. These transcripts are largely degraded by 6 hours of development [48], the second time point of the dataset.

For the absolute detection of transcripts (presence or absence calls), the method we used [49] replaces all MM probe values by a threshold value which is based on the mean PM value (after gcRMA transformation) of probesets that are very likely to have absent target transcripts. This removes the influence of probe sequence affinity and results in better performance than the MAS 5 algorithm.

### ***Significance of trends in zebrafish development***

For the zebrafish microarray data we first used a randomization approach to assess the significance of the difference between two curves of median expression across development (for example median expression of duplicates vs. singletons, or of essential genes vs. genes with no reported phenotype). If the two groups contain  $n_1$  and  $n_2$  genes, we pooled all these genes and randomly separated them into two new groups of same sizes ( $n_1$  and  $n_2$ ). Then we calculated and recorded the difference between the two new curves of median expressions across development. After repeating this randomization 10,000 times, we could define 1‰ and 1% confidence intervals.

Second, we calculated the Spearman correlation between developmental time and the difference between two curves of median expression across development. Bonferroni correction was applied to correct for multiple testing, considering the 9 tests computed with this microarray data (Figure 1; Figure 2; Figure 5; Figure S1; Figure S3; Figure S5A-D):  $\alpha=0.05/9=0.0056$ .

### ***Clustering of microarray data***

In order to identify genes lowly or highly expressed in early development, we used the Fuzzy C-Means soft clustering method implemented in the Bioconductor package Mfuzz [50]. After a pre-filtering step (genes with  $sd < 0.5$  were removed), we ran the algorithm with the number of clusters set to  $c=4$ . This gave one cluster of genes lowly expressed across development (3641 probesets, 2261 Ensembl genes), one of genes highly expressed (2175 probesets, 1175 Ensembl genes), one of genes whose expression increased (1714 probesets, 1123 Ensembl genes) and one of genes whose expression decreased (3306 probesets, 2446 Ensembl genes; Figure 3).

### ***Mouse EST data***

EST (Expressed Sequence Tags) data were retrieved from BGEE (dataBase for Gene Expression Evolution, <http://bghee.unil.ch/>), a database comparing transcriptome data between species [51], including EST libraries from UniGene [52]. The mapping of UniGene clusters on Ensembl genes is taken from Ensembl (version 48) [46], where a percentage of identity of 90% is set as the minimum threshold to link an Ensembl gene with a UniGene cluster. Each library has been annotated manually to ontologies of anatomy and developmental stages, if it was obtained under non pathological conditions, with no treatment (“normal” gene expression). We considered a gene expressed at one time point in development if at least one EST was mapped to this gene at this time point. Thus, we could retrieve the number of genes expressed at each time point of mouse (*Mus musculus*) development. From this set we extracted two groups to compare (for example essential/non essential, or duplicates/singletons). As the total number of ESTs available at each time point is different, we use at each time point the ratio of the numbers of genes expressed in the two groups. We obtained similar results when we defined a gene as expressed if it had at least two ESTs mapped to it. Also, considering the ratio of the mean number of ESTs per gene at each stage, instead of the ratio of the number of genes expressed at each stage, gave similar results (not shown). We used data from 297 EST libraries, spanning 26 different developmental stages (from TS01 to TS27), corresponding to a total of 633,307 ESTs.

A weighted linear regression between developmental time and expression ratios was fit to the data, and a F-test was run to assess if the slope was significantly different from zero. Weights were the total number of genes expressed at each stage. Bonferroni correction was applied to correct for multiple testing, considering the 6 ratios tested with mouse EST data (Figure 4A-C; Figure 6; Figure 7; Figure S2):  $\alpha=0.05/6=0.0083$ . To test for an hourglass-like model, we adjusted a parabola (polynomial model of order 2), as in Hazkani-Covo *et al.* [3]. We used an ANOVA to estimate if the increase in fit to the data ( $r$ ) between the linear and parabola models was significant. The same Bonferroni correction was applied to the ANOVA. This test was never significant, providing no evidence for a maximum or a minimum of the ratio during development (Dataset S2).

### ***Zebrafish In-situ data***

*In-situ* hybridization expression data from ZFIN [16] were retrieved using BGEE [51]. We considered only stages with more than 1000 genes expressed, starting when maternal genes are largely degraded (6 hours post-fertilization [48]). We retrieved all genes with at least one report of expression by *in-situ* hybridization, at each time point of zebrafish development. From this set we extracted two groups (for example essential and non-annotated genes), and analyzed their ratio across development using the same methodology as with ESTs (see above).

### ***Rate of protein evolution***

The orthology relationships, and the values of  $d_N$  (rate of non-synonymous substitution per codon) and  $d_S$  (rate of synonymous substitution per codon) were obtained from Ensembl version 48 [46]. We retrieved zebrafish genes with one-to-one orthologs in *Tetraodon nigroviridis* and *Takifugu rubripes* (divergence time is ~32 MYA between the two pufferfish species and ~150 MYA with *Danio rerio* [53]). We downloaded the pairwise  $d_N$  and  $d_S$  between *Tetraodon* and *Takifugu*, calculated with codeml from the PAML package in the Ensembl pipeline (model=0, NSsites=0) [54]. Ensembl considers that  $d_S$  values are saturated when they reach a threshold which is  $2 * \text{median}(d_S)$ . See [http://www.ensembl.org/info/about/docs/compara/homology\\_method.html](http://www.ensembl.org/info/about/docs/compara/homology_method.html) for further details.

We selected a set of 4937 genes having  $d_N$ ,  $d_S$  and Affymetrix expression data. Among them 620 genes were strict singletons in fishes (loss of duplicates after 2R and after the fish-specific genome duplication). At each time point we performed the Spearman correlation between the  $d_N/d_S$  ratio and expression, following Davis *et al.* [1]. A  $t$ -statistic was used to assess if the correlation coefficient was different from 0.

For the analysis in mouse we retrieved pairwise  $d_N$  and  $d_S$  between human and mouse, for genes with one-to-one human orthologs (14,333 genes). We kept only the singletons for 2R genome duplication and separated the 25% with the highest  $d_N/d_S$  and the 25% with the lowest  $d_N/d_S$  (607 genes in each group). We then compared the expression across development of these two groups using EST data. Using the 10% highest and lowest  $d_N/d_S$  gave similar results (not shown).

## ***Genotypes and phenotypes***

### *Zebrafish mutants*

Data on zebrafish mutants were retrieved from the Zebrafish Information Network ([http://zfin.org/zf\\_info/downloads.html](http://zfin.org/zf_info/downloads.html), April 2008) [16]. We selected mutant genotypes having a lethal or abnormal phenotype from the file “phenotype.txt”, paying attention that they were grown in normal conditions (ZDB-EXP-041102-1). These genotypes were mapped to ZFIN gene IDs using the file “genotype\_features.txt” and then to Affymetrix probesets using Biomart [55]. This resulted in a dataset of 252 ZFIN IDs associated with a lethal phenotype (79 Affymetrix probesets), and 2870 ZFIN IDs associated with an abnormal phenotype (461 probesets). Annotated normal phenotype data are rare in ZFIN, due to a lack of report of such mutants in the literature, so we used non-annotated as a reference (7246 ZFIN gene IDs with expression data).

To be sure that the technique used in the phenotype screen did not bias our analysis, we separated the dataset of genotypes having an abnormal phenotype by technique (file “genotype\_features.txt”): inversion, transgenic insertion, deficiency, point mutation, translocation, insertion, sequence variant or unspecified. Only transgenic insertions, point mutations and sequence variants provide enough data, with 343, 221 and 2424 ZFIN IDs respectively, corresponding to 309, 171 and 88 Affymetrix probesets respectively (Text S1 and Dataset S1).

### *Zebrafish morpholinos*

The morpholinos knock-down phenotypes were downloaded from ZFIN ([http://zfin.org/zf\\_info/downloads.html](http://zfin.org/zf_info/downloads.html), April 2008) [16]. We selected morpholinos (file “pheno\_environment.txt”) giving lethal or abnormal phenotypes (file “phenotype.txt”), paying attention that the genotypes were wild type (file “wildtypes.txt”). The probes were mapped to ZFIN gene IDs using the file “Morpholinos.txt” and then to Affymetrix probesets using Biomart [55]. Only “abnormal” phenotypes provided enough data, with 601 ZFIN IDs corresponding to 256 Affymetrix probesets (Text S1 and Dataset S1).

### *Mouse knock-outs*

Data on mouse mutants were retrieved from the Mouse Genome Database (<ftp://ftp.informatics.jax.org/pub/reports/index.html>, April 2008) [34]. We extracted

from the file MRK\_Ensembl\_Pheno.rpt all mutant genotypes having an annotated lethal (lethality-embryonic/perinatal, MP:0005374 and lethality-postnatal, MP:0005373), abnormal (other phenotypes detected) or normal phenotype (no phenotype detected, MP:0002873), and their mapping to Ensembl genes. We filtered on the technique used and kept only the mutants obtained with a targeted knock-out. Because different investigators do not report the same phenotypes for the same genes, we removed from the analysis all genes annotated to more than one group. We obtained 50 essential Ensembl genes (lethal phenotype), 164 non essential (normal phenotype), and 1939 whose loss of function is annotated abnormal (Dataset S2). Including genes annotated to more than one group, the group sizes were 1659, 564 and 3721 respectively, and the results were similar (not shown).

### ***Identification of duplicate genes***

Gene families were obtained from the HomolEns database version 3 (<http://pbil.univ-lyon1.fr/databases/homolens.html>), which is based on Ensembl release 41 [46]. HomolEns is build on the same model as Hovergen [56], with genes organized in families, which include pre-calculated alignments and phylogenies. In HomolEns version 3, alignments are computed with MUSCLE [57] (with default parameters), and phylogenetic trees with PhyML [58]. Phylogenies are computed on conserved blocks of the alignments selected with GBLOCKS [59]. Using the TreePattern functionality of the FamFetch client for HomolEns, which allows scanning for gene tree topologies [60], we selected sets of genes with or without duplications on specific branches of the vertebrate phylogenetic tree.

Regarding the fish-specific whole genome duplication, we found 1772 Ensembl IDs for duplicates in zebrafish, 8821 for singletons in zebrafish, 755 mouse orthologs of these duplicates, and 6843 mouse orthologs of these singletons. For the 2R whole genome duplications, we found 986 duplicates and 1266 singletons in zebrafish, and 2448 duplicates and 2705 singletons in mouse (Datasets S1 and S2).

### ***Gene Ontology Analysis***

Over and under representation of GO terms [25] was tested by means of a Fisher exact test, using the Bioconductor package topGO version 1.8.1 [61]. The reference set was all Ensembl genes mapped to a probeset of the zebrafish Affymetrix chip. The “elim”

algorithm of topGO was used, allowing to decorrelate the graph structure of the gene ontology, reducing non-independence problems. A False Discovery Rate correction was applied, and gene ontology categories with a FDR < 15% were reported.

### ***Tools***

R was used for statistical analysis and plotting (<http://www.R-project.org/>) [62], in conjunction with Bioconductor packages (<http://www.bioconductor.org/>, version 2.2)[45]. To retrieve genomic information we used the BioMart tool [55] or connected to the Ensembl MySQL public database [46].

## **3.5 Acknowledgements**

We thank Geisler R, Konantz M, Otto GW, Saric M, and Weiler C for making their microarray data publicly available. We thank Jérôme Goudet, Linda Z. Holland, Liliane Michalik and members of the MRR lab for helpful discussions.

## **3.6 Supporting Information**

Supporting material can be downloaded from:

<http://www.plosgenetics.org/article/info:doi/10.1371/journal.pgen.1000311#s5>

*Dataset S1:* Details and characteristics of zebrafish gene sets used in this study. FSGD: Fish Specific whole Genome Duplication.

*Dataset S2:* Details and characteristics of mouse gene sets used in this study. FSGD: Fish Specific whole Genome Duplication.

*Text S1:* Supplementary text.

*Figure S1:* Expression in zebrafish development of genes according to retention after vertebrate 2R whole genome duplications. Median expression profiles of vertebrate specific 2R duplicates in zebrafish in red dashed line and triangles, and of singletons in black solid line and circles. Legend as in Figure 2.

*Figure S2:* Variation across mouse development of the ratio of expressed vertebrate 2R singletons, relative to duplicates. Legend as in Figure 4.

*Figure S3:* Variation across zebrafish development of the Spearman correlation between gene sequence evolution and expression. Only singletons genes (for 2R and fish-specific genome duplications) were considered. We used the ratio of the rate of non-synonymous substitutions on the rate of synonymous substitutions ( $d_N/d_S$ ) as a measure of selective pressure. Correlations below the dashed line are significantly different from 0 (p-value < 0.05). The x-axis is in logarithmic scale. A gray box on the x-axis indicates the phylotypic period.

*Figure S4:* Expression in zebrafish development of genes with abnormal mutant phenotypes. Median expression profiles of zebrafish genes inducing abnormal phenotypes when non functional, for 4 different techniques, compared to non-annotated genes in black solid line and circles. The techniques are: morpholinos in purple dashed-dotted line and squares; transgenic insertions in green dashed line and triangles; point mutations in blue dashed line and diamonds; sequence variants in red dotted line and crosses. Points significantly different from the reference curve (non annotated genes) are filled. See Figure S5 for confidence intervals of the difference with the reference curve. The x-axis is in logarithmic scale. A gray box on the x-axis indicates the phylotypic period.

*Figure S5:* Significance of the expression difference between zebrafish genes inducing abnormal phenotypes when non functional and non-annotated genes for 4 different techniques. These randomization plots refer to Figure S4. Legend as in Figure 2B.

*Table S1:* Gene Ontology analysis. The two groups analyzed are the genes experiencing an increase of expression along development (late expression, cluster 4) and the genes experiencing a decrease of expression (early expression, cluster 3, Figure 3). Molecular Function and Biological process ontologies were analyzed with the “elim” algorithm of the Bioconductor package topGO (see Methods).

### 3.7 References

1. Davis JC, Brandman O, Petrov DA (2005) Protein evolution in the context of *Drosophila* development. *J Mol Evol* 60: 774-785.
2. Castillo-Davis CI, Hartl DL (2002) Genome Evolution and Developmental Constraint in *Caenorhabditis elegans*. *Mol Biol Evol* 19: 728-735.
3. Hazkani-Covo E, Wool D, Graur D (2005) In search of the vertebrate phylotypic stage: a molecular examination of the developmental hourglass model and von Baer's third law. *J Exp Zool B Mol Dev Evol* 304: 150-158.
4. Irie N, Sehara-Fujisawa A (2007) The vertebrate phylotypic stage and an early bilaterian-related stage in mouse embryogenesis defined by genomic information. *BMC Biology* 5: 1.
5. Hanada K, Shiu S-H, Li W-H (2007) The Nonsynonymous/Synonymous Substitution Rate Ratio versus the Radical/Conservative Replacement Rate Ratio in the Evolution of Mammalian Genes. *Mol Biol Evol* 24: 2235-2241.
6. Yang J, Li WH (2004) Developmental constraint on gene duplicability in fruit flies and nematodes. *Gene* 340: 237-240.
7. Holland LZ (2007) Developmental biology: A chordate with a difference. *Nature* 447: 153-155.
8. von Baer KE (1828) *Ueber Entwicklungsgeschichte der Thiere: Beobachtung und Reflexion*. Königsberg: Bornträger. 271 p.
9. Haeckel E (1874) *Anthropogenie oder Entwicklungsgeschichte des Menschen*. Leipzig: Engelmann. 732 p.
10. His W (1874) *Unsere Körperform und das physiologische Problem ihrer Entstehung*. Leipzig: FCW. Vogel. 224 p.
11. Gould SJ (1977) *Ontogeny and phylogeny*. Cambridge, Mass.: The Belknap Press of Harvard University Press. 501 p.
12. Duboule D (1994) Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. *Dev Suppl*: 135-142.
13. Raff RA (1996) *The shape of life: genes, development, and the evolution of animal form*. Chicago; London: University of Chicago Press. 520 p.
14. Sanetra M, Begemann G, Becker MB, Meyer A (2005) Conservation and co-option in developmental programmes: the importance of homology relationships. *Front Zool* 2: 15.
15. Irmeler I, Schmidt K, Starck JM (2004) Developmental variability during early embryonic development of zebra fish, *Danio rerio*. *J Exp Zool B Mol Dev Evol* 302: 446-457.
16. Sprague J, Bayraktaroglu L, Clements D, Conlin T, Fashena D, et al. (2006) The Zebrafish Information Network: the zebrafish model organism database. *Nucleic Acids Res* 34: D581-585.
17. Poe S, Wake MH (2004) Quantitative tests of general models for the evolution of development. *Am Nat* 164: 415-422.
18. Blalock EM, editor (2003) *A Beginner's Guide to Microarrays*: Springer. 368 p.
19. Jaillon O, Aury J-M, Brunet F, Petit J-L, Stange-Thomann N, et al. (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* 431: 946-957.



20. Brunet FG, Crollius HR, Paris M, Aury JM, Gibert P, et al. (2006) Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol Biol Evol* 23: 1808-1816.
21. Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, et al. (2008) The amphioxus genome and the evolution of the chordate karyotype. *Nature* 453: 1064-1071.
22. Chain FJ, Ilieva D, Evans BJ (2008) Duplicate gene evolution and expression in the wake of vertebrate allopolyploidization. *BMC Evol Biol* 8: 43.
23. Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH (2005) Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci USA* 102: 14338-14343.
24. Drummond DA, Raval A, Wilke CO (2006) A single determinant dominates the rate of yeast protein evolution. *Mol Biol Evol* 23: 327-337.
25. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25: 25-29.
26. Farre D, Bellora N, Mularoni L, Messeguer X, Alba MM (2007) Housekeeping genes tend to show reduced upstream sequence conservation. *Genome Biology* 8: R140.
27. Poe S (2006) Test of Von Baer's law of the conservation of early development. *Evolution* 60: 2239-2245.
28. Raff RA (1996) *The shape of life : genes, development, and the evolution of animal form*. Chicago ; London: University of Chicago Press. 520 p.
29. Bininda-Emonds OR, Jeffery JE, Richardson MK (2003) Inverting the hourglass: quantitative evidence against the phylotypic stage in vertebrate development. *Proc Biol Sci* 270: 341-346.
30. Holland LZ, Gibson-Brown JJ (2003) The *Ciona intestinalis* genome: when the constraints are off. *Bioessays* 25: 529-532.
31. Galis F, Metz JA (2001) Testing the vulnerability of the phylotypic stage: on modularity and evolutionary conservation. *J Exp Zool* 291: 195-204.
32. Wapinski I, Pfeffer A, Friedman N, Regev A (2007) Natural history and evolutionary principles of gene duplication in fungi. *Nature* 449: 54-61.
33. Richardson MK (1999) Vertebrate evolution: the developmental origins of adult variation. *Bioessays* 21: 604-613.
34. Eppig JT, Bult CJ, Kadin JA, Richardson JE, Blake JA, et al. (2005) The Mouse Genome Database (MGD): from genes to mice--a community resource for mouse biology. *Nucleic Acids Res* 33: D471-475.
35. Arthur W (2000) The concept of developmental reprogramming and the quest for an inclusive theory of evolutionary mechanisms. *Evol Dev* 2: 49-57.
36. Prud'homme B, Gompel N, Carroll SB (2007) Colloquium Papers: Emerging principles of regulatory evolution. *Proc Natl Acad Sci USA* 104: 8605-8612.
37. Koutsos AC, Blass C, Meister S, Schmidt S, MacCallum RM, et al. (2007) Life cycle transcriptome of the malaria mosquito *Anopheles gambiae* and comparison with the fruitfly *Drosophila melanogaster*. *Proc Natl Acad Sci USA* 104: 11304-11309.
38. Hinman VF, Nguyen AT, Cameron RA, Davidson EH (2003) Developmental gene regulatory network architecture across 500 million years of echinoderm evolution. *Proc Natl Acad Sci USA* 100: 13356-13361.
39. Nei M (2007) The new mutation theory of phenotypic evolution. *Proc Natl Acad Sci USA*.
40. Davidson EH, Erwin DH (2006) Gene regulatory networks and the evolution of animal body plans. *Science* 311: 796-800.

41. Solnica-Krezel L (2005) Conserved patterns of cell movements during vertebrate gastrulation. *Curr Biol* 15: R213-228.
42. Stern CD, editor (2004) *Gastrulation: From Cells to Embryo*: Cold Spring Harbor Laboratory Press.
43. Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, et al. (2007) ArrayExpress--a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res* 35: D747-750.
44. Wu Z, Irizarry R, A., Gentleman R, Martinez-Murillo F, Spencer F (2004) A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. *Journal of the American Statistical Association* 99: 909-917.
45. Gentleman R, Carey V, Bates D, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* 5: R80.
46. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, et al. (2007) Ensembl 2007. *Nucleic Acids Res* 35: D610-617.
47. Pelegri F (2003) Maternal factors in zebrafish development. *Dev Dyn* 228: 535-554.
48. Mathavan S, Lee SG, Mak A, Miller LD, Murthy KR, et al. (2005) Transcriptome analysis of zebrafish embryogenesis using microarrays. *PLoS Genet* 1: 260-276.
49. Schuster EF, Blanc E, Partridge L, Thornton JM (2007) Correcting for sequence biases in present/absent calls. *Genome Biol* 8: R125.
50. Futschik ME, Carlisle B (2005) Noise-robust soft clustering of gene expression time-course data. *J Bioinform Comput Biol* 3: 965-988.
51. Bastian F, Parmentier G, Roux J, Moretti S, Laudet V, et al. (2008) Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species. In: Heidelberg SB, editor. *Data Integration in the Life Sciences*. pp. 124-131.
52. Pontius JU, Wagner L, Schuler GD (2003) 21. UniGene: A Unified View of the Transcriptome. In: McEntyre J, Ostell J, editors. *The NCBI Handbook*. Bethesda, MD: National Library of Medicine (US), NCBI.
53. Benton MJ, Donoghue PC (2007) Paleontological evidence to date the tree of life. *Mol Biol Evol* 24: 26-53.
54. Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13: 555-556.
55. Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, et al. (2004) EnsMart: a generic system for fast and flexible access to biological data. *Genome Res* 14: 160-169.
56. Duret L, Mouchiroud D, Gouy M (1994) HOVERGEN: a database of homologous vertebrate genes. *Nucleic Acids Res* 22: 2360-2365.
57. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32: 1792-1797.
58. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52: 696-704.
59. Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 17: 540-552.
60. Dufayard J-F, Duret L, Penel S, Gouy M, Reichenmann F, et al. (2005) Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics* 21: 2596-2603.
61. Alexa A, Rahnenfuhrer J, Lengauer T (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* 22: 1600-1607.

62. R Development Core Team (2007) R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.



# 4 Molecular signaling in zebrafish development and the vertebrate phylotypic period

---

Aurélie Comte, Julien Roux, Marc Robinson-Rechavi

*This article is the result of the master project of Aurélie Comte. I suggested the subject and supervised it with Marc Robinson-Rechavi.*

## Abstract

During development vertebrate embryos pass through a stage where their morphology is most conserved between species, the phylotypic period (approximately the pharyngula). To explain the resistance to evolutionary changes of this period, one hypothesis suggests that it is characterized by a high level of interactions. Based on this hypothesis, we examined protein-protein interactions, signal transduction cascades and miRNAs over the course of zebrafish development, and the conservation of expression of these genes in mouse development. We also investigated the characteristics of genes highly expressed before or during the presumed phylotypic period. We show that while there is a high diversity of interactions during the phylotypic period (protein-DNA, RNA-RNA, cell-cell and between tissues), which is well conserved with mouse, there is no clear difference with later, more morphologically divergent, stages. We propose that the phylotypic period may rather be the expression at the morphological level of strong conservation of molecular processes earlier in development.

This article was published in *Evolution and Development* (2010) 12(2): 144-156

[doi: 10.1111/j.1525-142X.2010.00400.x](https://doi.org/10.1111/j.1525-142X.2010.00400.x)

## 4.1 Introduction

During metazoan embryonic development, the complexity of the organism increases from one cell to an integrated multi-cellular animal. This is accompanied not only by an increasing number of parts, but also by changes in the pattern of interactions among these parts [1]. In very early development, connections are limited, with the embryo mainly organized along two axes. When organ primordia form, the body becomes partitioned in "modules", between which numerous interactions take place. At late stages the organs continue to differentiate, but the "modules" are now semi-independent, and the interactions mainly occur within them. This model has been linked to the observation that mid-development is the most morphologically conserved period among vertebrate embryos [1,2,3,4,5], hence the term "phylotypic stage" or "phylotypic period".

In practice, such interactions must involve molecular pathways of signaling and regulation. Morphological models do not specifically predict that molecular pathways themselves should vary. But if signaling is dramatically different between early, middle ("phylotypic"), and late development, we expect to see changes in the activity of signaling pathways during development. Moreover, if changes in signaling are causal to the phylotypic period, we expect the timing of some changes in signaling to correspond with the boundaries of this period. Characterizing such molecular variation might help to reconcile divergent observations of developmental variation at the morphological and the genomic level [2,6,7,8,9].

In this study, we use expression information to relate zebrafish genes to developmental stages, and investigate the variation in protein-protein interactions, signal transduction cascades, and microRNA signaling. We also investigate whether the timing of gene expression is conserved in mouse. This allows us to distinguish signaling pathways which are most active in early, mid or late development, and can be related to the different phases of morphological integration.

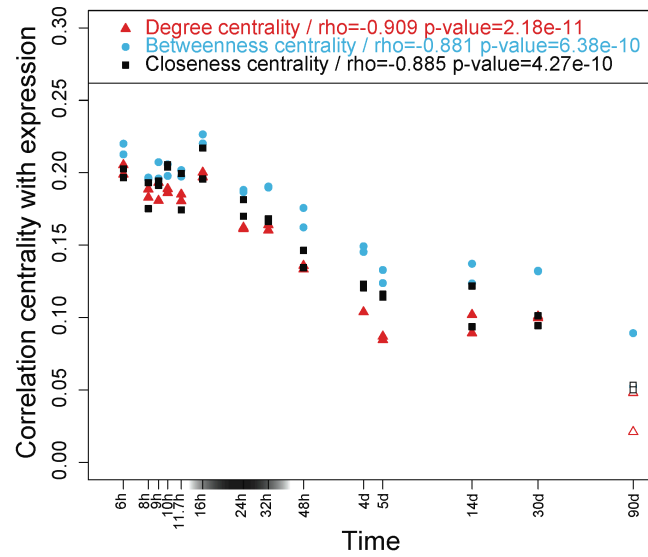


Figure 1: Variation of centrality in the protein-protein interaction network during development. Variation of the correlation between centrality and gene expression level with timing of gene expression across zebrafish development. The three curves represent degree centrality (red triangles), betweenness centrality (blue circles) and closeness centrality (black squares). Filled points indicate a significant correlation with expression at a given stage. Spearman correlations (coefficient  $\rho$ ) were computed between the correlation of centrality and expression, and developmental time. The gray box on the x-axis indicates the presumed phylotypic period. The x-axis is in logarithmic scale.

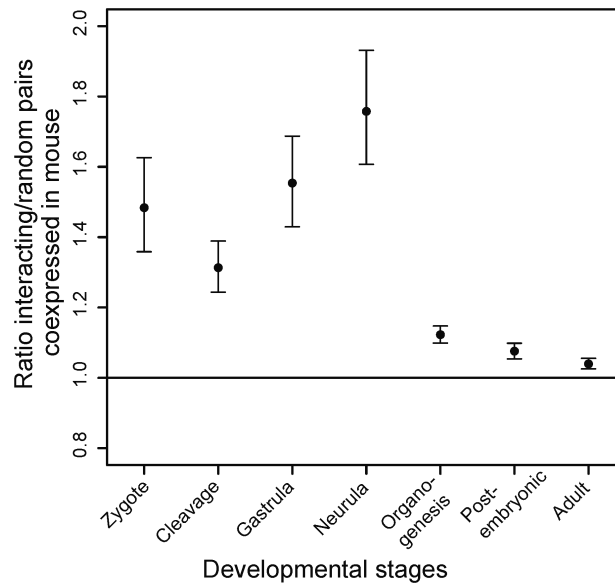


Figure 2: Conservation of co-expression of pairs of interacting proteins between zebrafish and mouse. Mean ratios of the number of pairs of interacting proteins whose co-expression is conserved between zebrafish and mouse at a given developmental meta-stage, to the number of random pairs of proteins whose co-expression is conserved between zebrafish and mouse. Bars represent percentiles of ratios (1% and 99% of repetitions). Organogenesis includes the presumed phylotypic period. The x-axis is not proportional to time, as the mapping of the stages of the two species compared on meta-stages is different. The horizontal line indicates a ratio of 1, i.e. conservation of interacting pairs not different from random pairs.

## 4.2 Results

### *Protein interconnectivity is highest in early development*

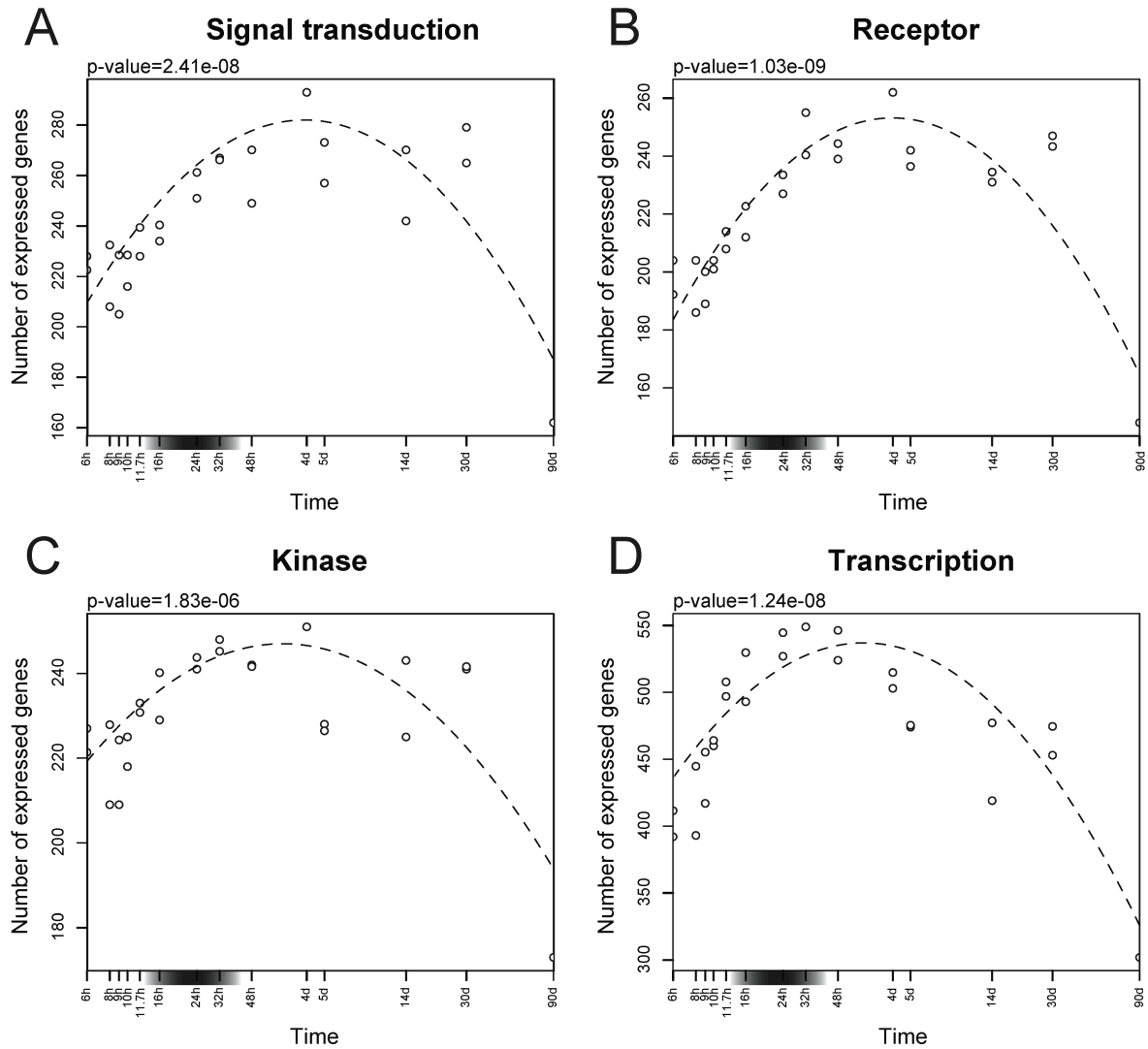
We first examined position in the protein-protein interaction (PPI) network, according to timing of expression of the genes encoding the interacting proteins. Proteins at the centre of the network are more connected than those at the network periphery. Consequently, determining the network centrality of a protein is equivalent to evaluating its level of connectivity. Of note, we transferred information on human interactions to the zebrafish; while this may affect the precision of our results, it is probable that trends are essentially correct [10].

We used three different measures to quantify the centrality of proteins: degree, betweenness and closeness centrality [11]. Degree centrality is defined as the number of links incident upon a node; it is a local measure. Betweenness and closeness centrality are global measures: the first reflects the number of occurrences of a node on shortest paths between other nodes, while the second reflects “shallowness” to other nodes. At each stage we computed Spearman’s correlation between these centrality measures and gene expression from microarray data, to remove the possible confounding effect of expression level on studies of connectivity [12]. The three centrality measures give similar results (Figure 1). At all stages the correlation is positive, confirming that highly expressed proteins tend to be central and to participate in many interactions. The correlation decreases over developmental time, suggesting that early expression has a higher relation to protein-protein connectivity than late expression. This is coherent with results from Liang and Li [13], who contrasted the centrality and connectivity of developmental vs. non-developmental genes. The presumed phylotypic period does not show any specific trend.

To verify the evolutionary relevance of these observations, we measured whether the orthologs of pairs of genes, which are both expressed in the same broad developmental stage in zebrafish, are also both expressed in the corresponding stage in mouse. While genes encoding pairs of interacting proteins have more conservation of co-expression than other genes at all stages, conservation is strongest in early development (zygote -



neurula, Figure 2). In later development, including the phylotypic period (included in organogenesis), the conserved co-expression of interacting proteins is much weaker.



*Figure 3: Variation of gene expression for signal transduction genes during development. Number of expressed genes per developmental stage annotated with GO terms containing (A) both "signal" and "transduction", (B) "receptor", (C) "kinase" and (D) "transcription". A polynomial model was fitted to the data (dashed line parabolas) with p-values indicated above each plot. The gray boxes on the x-axes indicate the presumed phylotypic period. The x-axes are in logarithmic scale.*

### ***Signal transduction is highest in the larva***

To investigate interactions between cells or tissues, we studied the expression of genes annotated with GO terms containing both "signal" and "transduction", as well as genes annotated as key components of signaling: receptors, kinases, and transcription genes. Each of these categories individually shares the general pattern of high correlation between PPI centrality and expression level early in development (Figure S4; Figure S5). The number of signal transduction, receptor and kinase genes expressed increases progressively to reach a maximum at 4 days (larval stage) and then decreases at later stages (Figure 3A, B, C). Excluding photoreceptors from the analysis of receptors, to check for potential bias due to eye development, does not modify observed trends (data not shown). Pairwise comparisons confirm that a significantly higher proportion of genes is expressed at 4 days than at 24h for signal transduction and receptors (comparison of proportions over both repetitions of the experiment, Bonferroni correction [5 tests]; signal transduction  $p = 0.0011$ ; receptor  $p = 0.0080$ ). Transcription genes peak earlier (Figure 3D), at 32h, which corresponds to late pharyngula, the stage most often associated with the phylotypic period [5]. There are significantly more transcription genes expressed at 24h or 32h than at 4 days (32h vs. 4 days:  $p = 0.0011$ ). But abundant expression remains during larval development. Genes which possess both transcription and receptor functions (i.e. nuclear receptors) show the same behavior as receptors (data not shown).

For all components of signaling tested, the expression of orthologs is significantly conserved in mouse development at all late stages, from organogenesis to adulthood (Figure 4); but not in early development. There is no specific peak of conservation in organogenesis, which includes the phylotypic stage.

Thus signal transduction appears important, and evolutionarily conserved, over a large period of development, which starts during the phylotypic period but lasts into post-embryonic development.

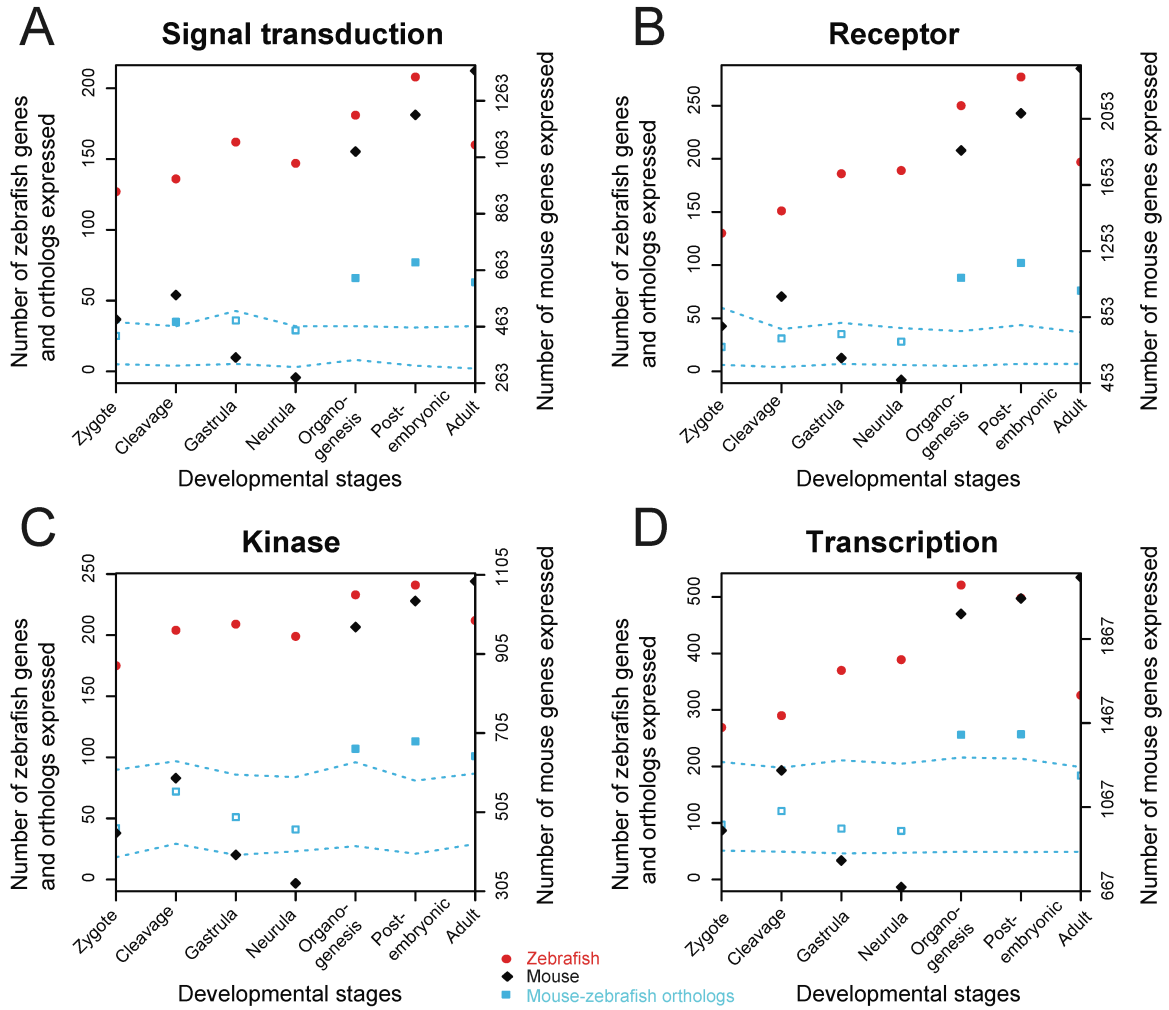


Figure 4: Conservation of gene expression for signal transduction genes between zebrafish and mouse. Number of zebrafish (red circles) and mouse (black diamonds) genes, and ortholog pairs (blue squares) expressed per developmental stage for: (A) signal transduction, (B) receptors, (C) kinases and (D) transcription. The dotted lines represent the 1% confidence interval for conserved expression of orthologs; significant numbers of orthologs expressed are represented by filled squares. Organogenesis includes the presumed phylotypic period. The x-axis is not proportional to time, as the mapping of the stages of the two species compared on meta-stages is different. The scale of the y-axis is different for mouse, as more data are available.

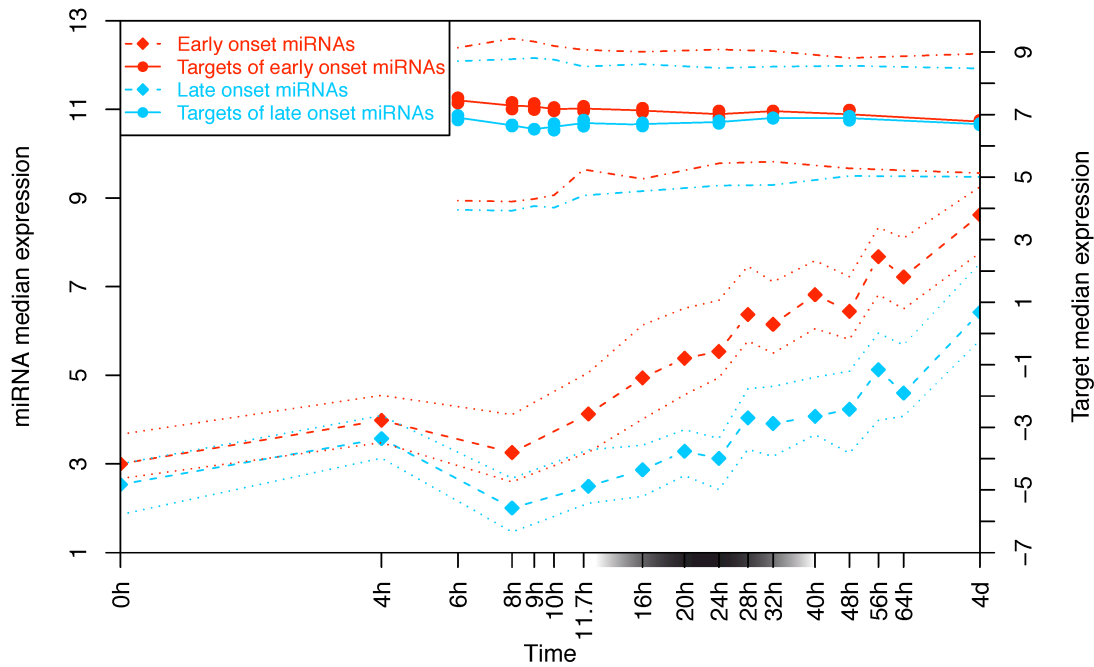


Figure 5: Variation of miRNA and target genes expression during development. Median expression of “early onset” miRNAs (red dashed line, diamonds;  $n=65$ ) and their targets (red line, circles;  $n=119$ ), and of “late onset” miRNAs (blue dashed line, diamonds;  $n=44$ ) and their targets (blue line, circles;  $n=253$ ). Dotted lines represent quartiles of miRNA expression; dot-dashed lines represent quartiles of target gene expression. Differences between the two target groups and significance are show in Figure S3. The gray box on the x-axis indicates the presumed phylotypic period. The x-axis is in logarithmic scale.

### ***miRNA expression increases progressively through development***

It has been proposed that the control of protein coding genes by miRNAs leads to a gain of developmental precision at the cost of a loss of evolutionary plasticity [14]. This suggests that the less morphologically variable developmental stages could be under stronger miRNA control.

The expression of miRNAs during zebrafish development (Figure S2) suggests a classification into two categories: “early onset” miRNAs whose expression starts to increase before the presumed phylotypic period (11.7h, segmentation), and “late onset” miRNAs whose expression rises later (28h, pharyngula; Figure 5). In both groups a peak of expression is detected at 4h (blastula). It corresponds most probably to the maternal-zygotic transition [15]. No other peak of expression is noticed along development.

Expression of targets of the “late onset” is stable across development, while “early onset” targets experience a small decrease during development (Figure 5, Figure S3). As miRNAs are negative regulators of gene expression, the observation of a decrease in the expression level of targets of “early onset” miRNAs once these miRNAs are expressed is not surprising. However the interpretations of this result should be considered with care. The difference in median expression between the targets of the two categories of miRNAs is globally not significant across development, as assessed by a randomization (except for one of the replicates at time point 9h; Figure S3). It is probable that by using gene and miRNA expression data from the whole organism, we have missed fine regulation in specific regions of the embryo. It is also possible that the high rate of false positives in databases of target predictions [16] renders this result less accurate or precise.

There is no comparable data on expression of miRNAs during development of other vertebrate species, so we cannot investigate evolutionary conservation of these patterns.

### ***Characteristics of genes expressed during different developmental periods***

As an alternative to studying the expression profile of groups of candidate genes, we used soft clustering of expression profiles to generate groups of genes, whose properties may be related to the patterns of evolution and development (Figure 6; Figure S1). This provided us with three sets of genes with interesting profiles in development: (i) Expression of the “early” genes is high early in development, and decreases to reach a stable low level by the presumed phylotypic period. (ii) Expression of the “organogenesis” genes is low at early stages, then increases strongly at the presumed phylotypic period and remains high during larval development, with a decrease in adults. (iii) Expression of the “late” genes is low both in early development and during the phylotypic period, with a later increase towards the larval stage.

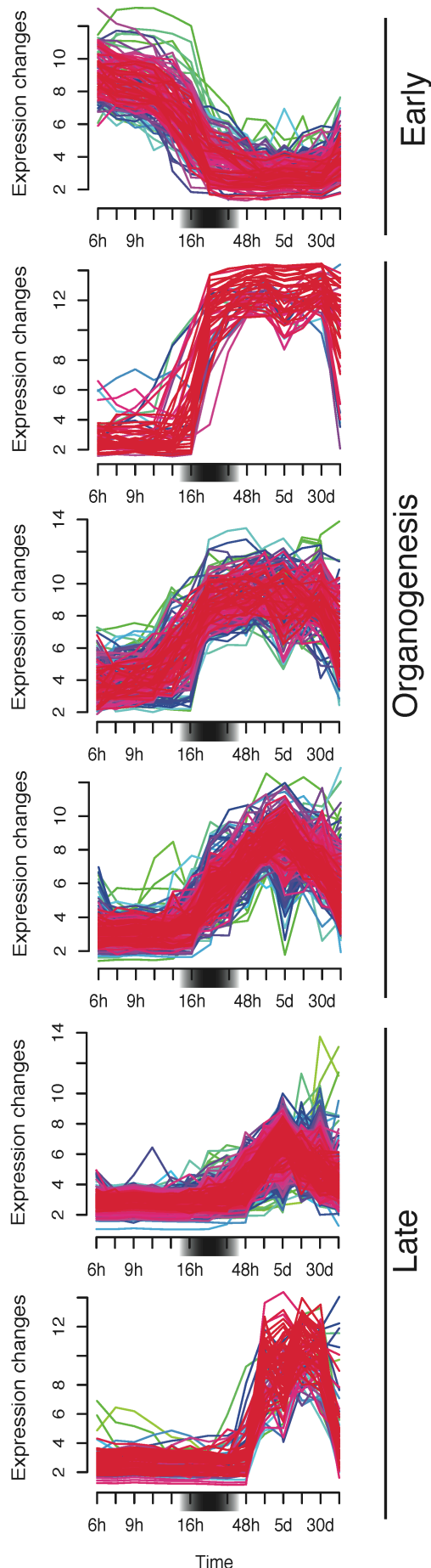
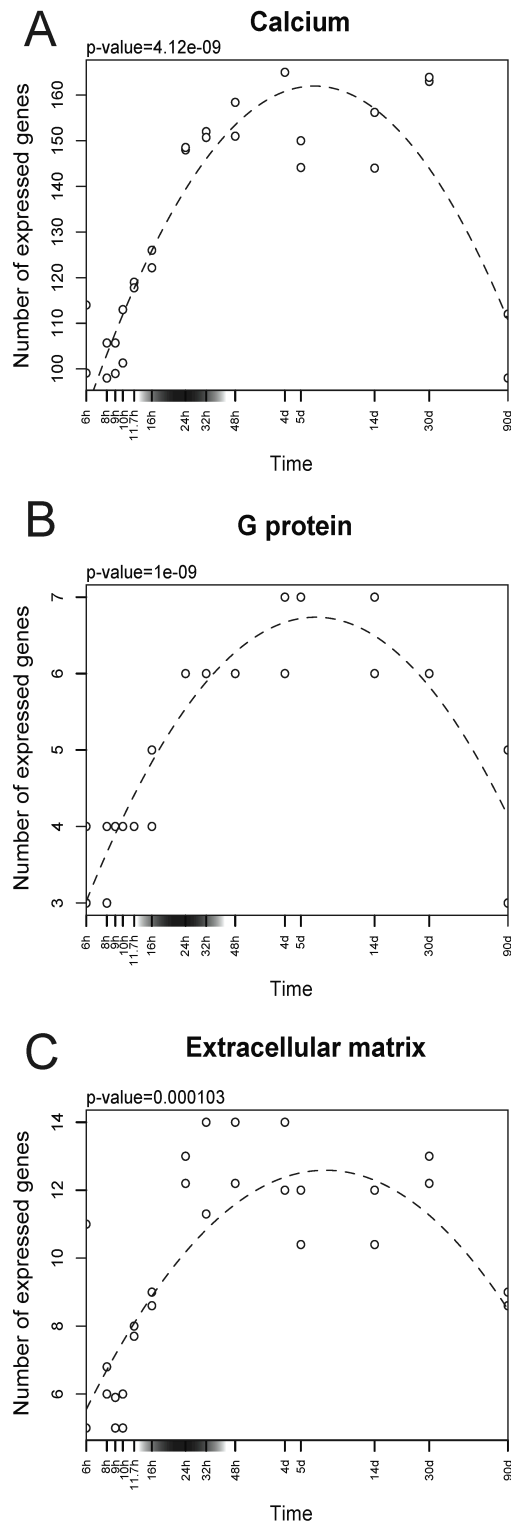


Figure 6: Expression profiles of “early”, “organogenesis” and “late” genes. Each line represents a gene, color coded according to how well it is represented by the cluster, from yellow or green for low membership scores, to red or purple for high membership scores. The gray boxes on the x-axes indicate the presumed phylotypic period. All 25 clusters are presented in Figure S1.

The average number of abnormal phenotypes reported for mutation of genes from these groups differs significantly ( $p = 0.0078$ , Kruskal-Wallis test). Mutation of “early” genes results in the most abnormal phenotypes (average of 10.5 vs. 5.28 for “organogenesis” genes and 6.86 for “late” genes). There is also a significant difference between the three categories for the number of anatomical structures in which each gene is detected ( $p = 5.85e-11$ , Kruskal-Wallis test). This is mostly due to “late” genes being expressed in fewer structures (5.48 vs. 10.3 for “organogenesis” genes and 9.5 for “early” genes); in other words, “late” genes are more tissue-specific. As might be expected, expression of “early” genes is enriched in presumptive structures. Expression of “organogenesis” genes is enriched in numerous anatomical structures, most of them related to the nervous system,

the visual system, the muscle, the heart and the pancreas. And expression of “late” genes is enriched in the visual, intestinal and nervous systems.

An analysis of GO terms (Table 1) shows notably that “organogenesis” genes are enriched in proteins localized in the extracellular matrix, and in heterotrimeric G-protein complexes. This suggests a role in mediating cell or tissue interactions. Also of interest, these genes are enriched in molecular functions and biological processes related to calcium; calcium is a secondary messenger in many signal transduction pathways. However, calcium also plays a role in muscle contraction, and terms related to muscle are also enriched in “organogenesis” genes. It is difficult with our data to distinguish these two roles of calcium in development. Looking at the global pattern of genes from these GO categories, they have a similar expression profile to the signal transduction genes, with highest expression in larva (Figure 7), and higher conservation of expression with mouse in organogenesis and post-embryonic development (Figure 8).



*Figure 7: Variation of gene expression for genes involved in signaling in organogenesis. Number of expressed genes per developmental stage for: (A) calcium (GO:0005262, GO:0019855 and GO:0005509; 196 genes); (B) heterotrimeric G protein complex (GO:0005578; 7 genes); (C) proteinaceous extracellular matrix (GO:0005834; 18 genes). A polynomial model was fitted to the data (dashed line parabola) with p-values indicated above each plot. The gray boxes on the x-axes indicate the presumed phylotypic period. The x-axes are in logarithmic scale.*



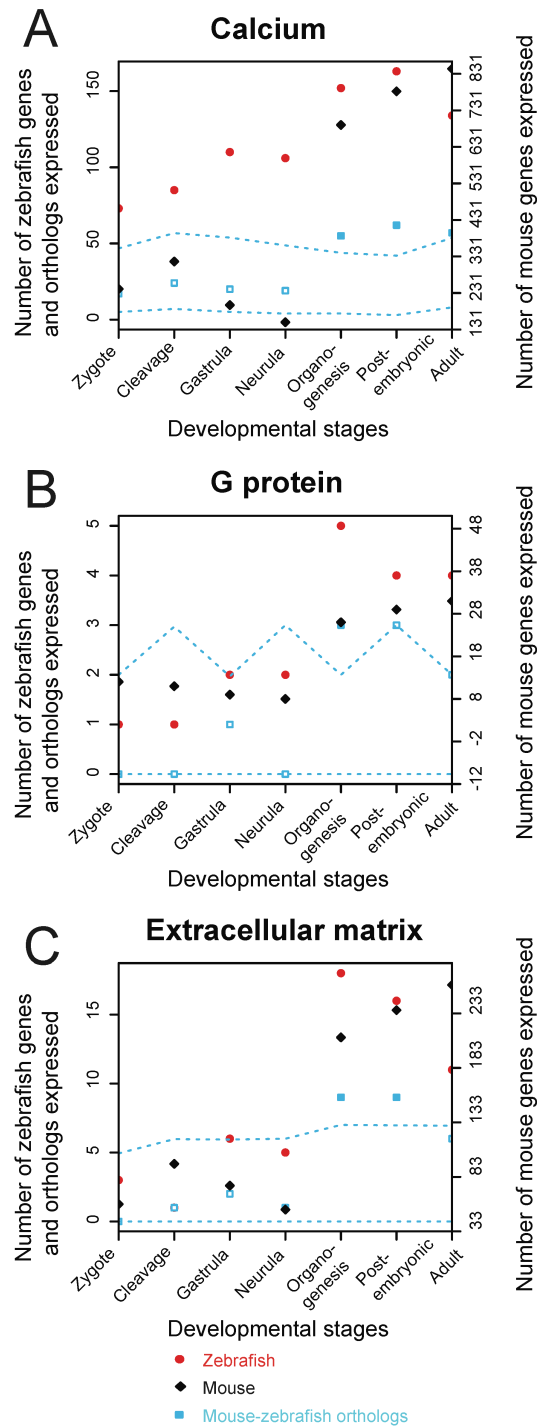


Figure 8: Conservation of gene expression for genes involved in signaling in organogenesis between zebrafish and mouse. Number of zebrafish (red circles) and mouse (black diamonds) genes, and ortholog pairs (blue squares) expressed per developmental stage for: (A) calcium (GO:0005262, GO:0019855 and GO:0005509; 174 zebrafish and 862 mouse genes, 71 orthologs); (B) heterotrimeric G protein complex (GO:0005578; 5 zebrafish and 31 mouse genes, 3 orthologs) and (C) proteinaceous extracellular matrix (GO:0005834; 20 zebrafish and 265 mouse genes, 12 orthologs). The dotted lines represent the 1% confidence interval for conserved expression of orthologs; significant numbers of orthologs expressed are represented by filled squares. Organogenesis includes the presumed phylotypic period. The x-axis is not proportional to time, as the mapping of the stages of the two species compared on meta-stages is different. The scale of the y-axis is different for mouse, as more data are available.

## 4.3 Discussion

On the basis of Raff's [1] hypothesis that the conserved morphology between vertebrate species at the phylotypic period could be the result of specific interactions, we investigated different molecular aspects related to interactions and signaling during zebrafish development. It should be noted that the data available do not allow us to test directly the hypothesis about differences in modularity between developmental stages. We can only evaluate the overall importance of molecular interactions and signaling, not whether it occurs inside or among "modules". But our working hypothesis is that major changes in signaling will probably affect the extent to which different regulatory mechanisms are used. Thus if the phylotypic period is defined by a specific pattern of interactions, we expect this period to be characterized by a specific signature of expression of genes involved in signaling and regulation.

A first notable observation is that many measures of signaling do present a peak during development (Figure 3; Figure 7), and that these peaks seem to be evolutionarily conserved since they are also detected in mouse (Figure 4; Figure 8). This stands in contrast to the monotonous decrease we previously reported for evolutionary constraints on the genome [9], and which is also observed for PPI centrality (Figure 1; Figure 2). The other notable observation is that the peak rarely corresponds to the morphologically defined phylotypic period.

The only feature which peaks close to the phylotypic period is the number of transcription genes expressed (Figure 3D). Combined with the onset of expression of a first wave of miRNAs (Figure 5), this could be seen as supportive of strong regulation of gene expression during this period. But these and other features which increase during the phylotypic period do not decrease until much later; most present maxima during larval development (Figure 3; Figure 7). There are for example more miRNAs expressed after than during the phylotypic stage, which is indicative of tight regulation of gene expression in late development. Moreover, when we classify genes according to their pattern of expression during development, there is no class of genes which peak specifically during the phylotypic period, but rather many genes which increase during that period, then do not decrease significantly until adulthood (Figure 6). These "organogenesis" genes are enriched in proteins with a potential role in signaling between cells or tissues, considering their cellular localization and their relation with calcium. In zebrafish, intracellular as well as localized and long range intercellular

calcium signaling patterns have been observed from cleavage to segmentation [17]. These calcium signaling events have been shown to be involved in dorso-ventral and left-right patterning, convergent extension during gastrulation and somite formation. A role for calcium signaling in development is not restricted to zebrafish, as experiments have also implicated calcium in dorso-ventral patterning and convergent extension movement as well as neural induction in *Xenopus*, in left-right patterning in mouse and chicken, and in somite formation in chicken [18,19]. Indeed, the expression of calcium signaling genes in organogenesis and larval stages is conserved between zebrafish and mouse (Figure 8A).

The late peak in the number of signal transduction and receptor genes expressed suggests a major role for cell, tissue, and receptor-ligand interactions. At the same time the majority of miRNAs are expressed at a high level and consequently mediate numerous RNA-RNA interactions. This probably reflects the increasing complexity of the organism, and the need for specific regulation in differentiated organs and tissues. This specialization is supported by the tissue specificity of “late” genes.

While the separation between a phylotypic period and further organogenesis and larval development is thus not clearly defined by any type of gene expression, early development does present a quite specific pattern. This can be seen e.g. in the conservation of gene co-expression between zebrafish and mouse: whereas the conservation of co-expression of interacting proteins is highest in early development (Figure 2), conservation of signaling gene expression is lowest (Figure 4). Moreover, we can identify a cluster of 160 genes that are highly expressed early in development, but have practically lost expression by pharyngula (24h), and remain at very low levels thereafter (Figure 6). These specific “early” genes are enriched in terms related to body plan specification (Table 1). Thus the information for the body plan appears to be laid out before the phylotypic period, when genes are under the strongest evolutionary constraints [9]. The observation that mutation of these “early” genes produces the most diverse abnormal phenotypes is also consistent with a key role for early development, rather than for the phylotypic period. These early genes appear to participate highly in conserved protein-protein interactions (Figure 1; Figure 2), whereas miRNA regulation is almost absent (Figure 5; [20]). This pattern is inversed from organogenesis to larval development (high miRNA regulation, small role of protein-protein interactions).

These results pose the question of why a phylotypic period is observed at the morphological level. True, there are many molecular interactions around that period of zebrafish development, and they seem to be conserved with mouse. But they mostly continue into further organogenesis and larval development, sometimes even reaching a maximum during the larval stage, which is not morphologically conserved. We suggest that a solution lies in realizing that morphology at each stage of development probably depends on an interaction between morphology at the previous stage and the genes expressed, which act to modify this morphology [21]. Under this simple assumption, early development would be constrained by its starting point, i.e. the very divergent zygotic morphologies [1,22]; under the influence of the conserved genetic determinants of early development [9], morphology should tend to converge ([also suggested for insects [23]]); and finally the rapidly evolving genes expressed in later development should cause a corresponding divergence in morphology. This explanation allows for a minimum in morphological divergence at mid development, without any corresponding peak in genetic or molecular processes.

## **4.4 Conclusion**

There are high levels of interactions between molecules, and between cells and tissues, during the presumed phylotypic period, conserved between zebrafish and mouse. But there does not appear to be a marked boundary in levels or types of interactions, nor in zebrafish-mouse conservation, between that period and later development, where morphology is more divergent between species. On the other hand, expression and interaction data show a marked change between early (pre-phylotypic period) and later development. Early expressed genes appear to be both more conserved between zebrafish and mouse, and regulated by different pathways, than other genes, with more protein-protein interactions and little or no miRNA regulation. We propose that morphological conservation at the phylotypic period is a consequence of this early genetic conservation.

## 4.5 Material & Methods

### *Microarray data and clustering*

Microarray data of zebrafish (*Danio Rerio*) development were retrieved from ArrayExpress (E-TABM-33; [24]). This experiment used an Affymetrix GeneChip Zebrafish Genome Array (A-AFFY-38) with 15,617 probes, which correspond to 8,922 Ensembl genes [25]. 15 stages, two replicates per time point, were sampled: 15min, 6, 8, 9, 10, 11.7, 16, 24, 32, 48 hours, 4, 5, 14, 30 and 90 days, spanning zygote, gastrula, segmentation, pharyngula, hatching, larval, juvenile and adult stages.

Raw CEL files were normalized using the gcRMA package [26] of Bioconductor [27]. We used the “affinities” model of gcRMA, which uses mismatch probes as negative control probes to estimate the non-specific binding of probe sequences. The normalized values of expression are in log2 scale, which attenuates the effect of outliers.

Presence and absence calls were retrieved from ArrayExpress. The method used for absolute detection of transcripts was the MAS5 algorithm.

For the 1,965 Ensembl genes that are represented by more than one probe, we used the mean of all the probe values as the gene expression value, and we considered the gene present if more than half of its probe calls determined it as present.

The two replicates were used for calculations and plotting except for clustering where we used the average of the two replicates. As in Roux and Robinson-Rechavi [9] we did not consider the first time point of the data (15min, fertilization).

The genes were separated in 25 clusters (Figure S1) using the fuzzy c-means soft clustering algorithm implemented in the Mfuzz package [28] of Bioconductor. From these clusters we formed three groups of genes: highly expressed in early development (cluster 15; 160 “early” genes), highly expressed at the presumed phylotypic period (clusters 1, 20 and 23; 475 “organogenesis” genes), and highly expressed after the presumed phylotypic period (clusters 3 and 8; 412 “late” genes).

### *Protein-protein interactions*

Human protein-protein interactions were downloaded from the BioGRID [29], IntAct [30] and HPRD [31] databases. Interacting proteins were respectively mapped from HGNC symbol, Uniprot Accession and EntrezGene ID to Ensembl human genes. 671 EntrezGene IDs that corresponded to more than one Ensembl human gene were

removed. The Ensembl human-zebrafish one-to-one orthologs were retrieved from Ensembl. We merged the interaction data of the three databases yielding a dataset of 5,277 protein pairs with associated expression data.

For each developmental stage we retained interactions for which both interacting proteins were expressed according to the present/absent calls of the microarray data.

Degree, betweenness and closeness centrality measures [11] were calculated for each interacting protein at each stage using the R igraph package (<http://www.R-project.org/>; [32]). Spearman correlation between gene expression and centrality measures was performed for each stage.

### ***Signal transduction genes***

Zebrafish genes and their associated GO IDs were retrieved with Biomart [33] and the GO terms were downloaded from Gene Ontology (Nov 3, 2008; [34]). Genes annotated with GO terms that contained “signal” and “transduction”, “receptor”, “kinase”, or “transcription” were retrieved. This resulted in 421 signal and transduction, 413 receptor, 299 kinase and 691 transcription genes for which expression data existed; 47 genes were annotated with both “receptor” and “transcription” terms (i.e. nuclear receptors). The numbers of expressed genes for each stage and each replicate were determined according to the present/absent calls of the microarray data; the mean of the two replicates was used.

A linear regression between developmental time and number of expressed genes was fit to the data. To test for an hourglass-like model, we adjusted a parabola (polynomial model of order 2), as in Roux and Robinson-Rechavi [9]. We used an ANOVA to estimate if the increase in fit to the data ( $r$ ) between the linear and parabola models was significant. A Bonferroni correction was applied to correct for multiple testing, considering the 7 regressions of Figure 3 and Figure 7.

### ***Gene Ontology analysis***

Over and under representation of GO terms for “early”, “organogenesis” and “late” genes were tested with a Fisher exact test using the Bioconductor package topGO [35]. The reference set was all the Ensembl genes that were represented by a probe on the microarray. The “elim” algorithm of topGO was used, allowing decorrelation of the GO

graph structure, reducing non-independence problems. A False Discovery Rate correction was applied and gene ontology terms with a FDR < 5% were reported.

### ***Phenotypes and localization of expression data***

Zebrafish genotypes and phenotypes were recovered from the Zebrafish Information Network (ZFIN; July 2008; [36]). We selected the phenotypes corresponding to single gene mutants grown in normal conditions and to wild-type lines treated with only one morpholino targeting a single gene. The localization of gene expression for wild-type lines grown in normal conditions was also retrieved from ZFIN. Genes were mapped from ZFIN IDs to Ensembl IDs; 573 ZFIN IDs that correspond to more than one Ensembl ID were removed. There was mutant phenotype information for 22 “early” genes, 29 “organogenesis” genes and 7 “late” genes. And 96 “early” genes, 294 “organogenesis” genes and 211 “late” genes had localization of expression data.

The significance of the difference between the mean numbers of abnormal phenotypes or structures with expression per gene of the three categories was determined with a Kruskal-Wallis test. When the difference was statistically significant, pairwise Wilcoxon tests were performed; p-values were adjusted for multiple testing using the Bonferroni correction.

Enrichment and depletion of expression in anatomical structures (ZFIN) for “early”, “organogenesis” and “late” genes were tested with a Fisher exact test using a version of the Bioconductor package topGO [35] modified to handle any OBO ontology (Alexa and Roux, unpublished). The reference set, the algorithm and the FDR value are the same as for the GO analysis. We used only structures that show expression of at least 5 genes.

### ***miRNAs targets and expression***

Zebrafish miRNAs were downloaded from the miRBase database [37].

A time series of miRNA microarray data during zebrafish development [20] was retrieved (GSE2625) from GEO [38]. In this experiment a microarray developed for the detection of mammalian miRNAs was used to measure the expression of zebrafish miRNAs, which is made possible by the very strong sequence conservation of miRNAs. 15 stages were sampled: 0, 4, 8, 12, 16, 20, 24, 28, 32, 40, 48, 56, 64 hours and 4 days, spanning zygote, blastula, gastrula, segmentation, pharyngula, hatching and larval stages, as well as male and female adults. Adult time points were removed from our

analyses, as their expression value did not correspond to what was reported in Wienholds et al. [20], even after normalization. Expression data was normalized using the control probes pre-3, pre-4 and pre-5, and subsequently log transformed. Each miRNA is represented by five probes on the microarray. We used the mean of all the probe values as the miRNA expression value. We thus had expression data for 109 zebrafish miRNAs.

The miRNAs were separated in 2 clusters (Figure S2) using the fuzzy c-means soft clustering algorithm implemented in the Mfuzz package [28] of Bioconductor. We defined the 65 miRNAs from cluster 1 as “early onset” and the 44 miRNAs from cluster 2 as “late onset”.

EIMMo [39] target predictions for zebrafish miRNAs were retrieved from <http://www.mirz.unibas.ch/miRNAtargetPredictionBulk.php> (v3, January 2009). Targets were mapped from RefSeq IDs to Ensembl zebrafish genes. Ensembl genes that corresponded to more than one RefSeq IDs were removed.

Among the genes for which we have expression data, 119 are targeted only by “early onset” miRNAs and 253 only by “late onset” miRNAs. To assess the significance of the difference between median expression across development of the “early onset” miRNAs targets and the “late onset” miRNAs targets, we used a randomization approach (as in [9]). We pooled all the targets, randomly formed two new groups of the same size as the original groups ( $n_1=119$ ,  $n_2=253$ ) and calculated the difference in median expression between the two random groups, with 10,000 repetitions.

### ***Conservation of gene expression in mouse***

Expression information (Affymetrix, “high quality”) during development was retrieved for zebrafish (6,305 genes) and mouse (*Mus musculus*; 17,192 genes) from Bgee, a database to compare expression data between species [40]. The Ensembl mouse-zebrafish one-to-one orthologs were retrieved from Ensembl. While homologous developmental stages cannot be defined precisely, Bgee implements broadly defined meta-stages which can be compared between species. A precise description of the meta-stages and the correspondence between mouse or zebrafish stages to them can be found in the files stages.obo and stage\_association.txt downloadable at <http://bgee.unil.ch/bgee/bgee?page=download>.



To quantify the conservation of co-expression of interacting proteins over developmental meta-stages, we calculated for each meta-stage the number of interacting pairs of proteins for which both zebrafish and mouse one-to-one orthologs are expressed. This was compared to the co-expression of random pairs of zebrafish genes (10,000 randomizations). We plot the mean ratios of observed co-expression of PPI pairs to random pairs.

Zebrafish and mouse genes and their associated GO IDs were retrieved with Biomart and the GO terms were downloaded from Gene Ontology (June 25, 2009). Genes annotated with GO terms that contained “signal” and “transduction”, “receptor”, “kinase”, or “transcription” were retrieved. We kept the mouse-zebrafish one-to-one orthologs with GO annotation and expression data in both species. This resulted in 98 pairs for signal transduction, 124 for receptor, 127 for kinase and 307 for transcription. We calculated the total number of mouse and zebrafish genes of each gene category expressed at each meta-stage, as well as the number of ortholog pairs both expressed at each meta-stage. To assess the significance of the number of orthologs expressed, we randomly created pairs of mouse-zebrafish genes from the same gene category. Repeating this process 10,000 times, we could define 1% confidence intervals.

## 4.6 Acknowledgments

We thank A. Reymond and B. Piasecka for helpful discussions, and two anonymous reviewers for helpful comments. We acknowledge funding from Etat de Vaud, Swiss National Science Foundation grant 116798, and the Swiss Institute for Bioinformatics.

## 4.7 Supporting Information

Supporting material can be downloaded from:

<http://www3.interscience.wiley.com/journal/123323445/supinfo>

*Figure S1:* Gene clustering according to expression in development. Twenty-five clusters of genes obtained by soft clustering. Cluster 15 corresponds to the “early” genes. Clusters 1, 20 and 23 correspond to the “organogenesis” genes. Clusters 3 and 8 correspond to the “late” genes. Soft clustering assigns a gene gradual degrees of membership to a cluster. The membership scores indicate how well the gene is represented by a cluster, and are color-coded from yellow or green for low membership

scores to red or purple for high membership scores. The gray boxes on the x-axes indicate the presumed phylotypic period.

*Figure S2:* miRNA clustering according to expression in development. Two clusters of miRNA obtained by soft clustering. Soft clustering assigns a miRNA gradual degrees of membership to a cluster. The membership scores indicate how well the miRNA is represented by a cluster, and are color-coded from yellow or green for low membership scores to red or purple for high membership scores. The gray boxes on the x-axes indicate the presumed phylotypic period.

*Figure S3:* Variation of miRNA target genes expression during development. Difference in median gene expression between targets of “early onset” and “late onset” miRNAs. The dashed lines represent the 5% confidence interval; significant differences are represented by filled circles. The gray box on the x-axis indicates the presumed phylotypic period. The x-axis is in logarithmic scale.

*Figure S4:* Variation of centrality in the protein-protein interaction network for signal transduction genes during development. Same as Figure 1, but restricted to the gene categories used in Figure 3.

*Figure S5:* Variation of centrality in the protein-protein interaction network for non-signal transduction genes during development. Same as Figure 1, but restricted to the genes that do not belong to any of the categories used in Figure 3 (n=7399).

## 4.8 Tables

Table 1: Gene Ontology terms enriched or depleted according to expression profile in development

Expression profile	GO <sup>a</sup>	Direction	GO ID	Term	Observed	Expected	p-value	Adjusted p-value (FDR)
Early	BP	Enriched	GO:0007368	Determination of left/right symmetry	9	0.41	6.40 E-11	4.90 E-08
			GO:0035050	Embryonic heart tube development	9	0.54	1.10 E-09	4.21 E-07
			GO:0007498	Mesoderm development	11	0.45	9.30 E-09	2.37 E-06
			GO:0001707	Mesoderm formation	5	0.25	2.40 E-06	4.60 E-04
			GO:0009953	Dorsal/ventral pattern formation	9	0.68	5.50 E-06	8.43 E-04
			GO:0030903	Notochord development	4	0.23	5.00 E-05	6.13 E-03
			GO:0040007	Growth	6	0.7	5.60 E-05	6.13 E-03
			GO:0042664	Negative regulation of endodermal cell fate specification	3	0.12	1.60 E-04	1.53 E-02
			GO:0001706	Endoderm formation	3	0.14	2.80 E-04	2.38 E-02
			GO:0009798	Axis specification	3	0.19	6.60 E-04	4.21 E-02
			GO:0045893	Positive regulation of transcription, DNA-dependent	3	0.19	6.60 E-04	4.21 E-02
			GO:0048264	Determination of ventral identity	3	0.19	6.60 E-04	4.21 E-02
MF	Enriched		GO:0003700	Transcription factor activity	19	6.24	8.60 E-06	3.21 E-03
			GO:0008083	Growth factor activity	6	0.65	3.50 E-05	6.53 E-03
			GO:0043565	Sequence-specific DNA binding	14	4.97	3.50 E-04	4.35 E-02
			GO:0005634	Nucleus	28	13.11	3.70 E-05	5.74 E-03
Organo-genesis	BP	Enriched	GO:0006816	Calcium ion transport	24	4.93	3.30 E-11	2.53 E-08
			GO:0006096	Glycolysis	9	1.44	6.20 E-06	2.37 E-03
			GO:0030239	Myofibril assembly	5	0.41	1.70 E-05	4.34 E-03
			GO:0015671	Oxygen transport	5	0.62	2.00 E-04	3.32 E-02

		GO:0051258	Protein polymerization	6	0.98	2.60 E-04	3.32 E-02
		GO:0006813	Potassium ion transport	6	0.98	2.60 E-04	3.32 E-02
MF	Enriched	GO:0019855	Calcium channel inhibitor activity	22	5.05	2.30 E-09	5.22 E-07
		GO:0005262	Calcium channel activity	22	5.11	2.80 E-09	5.22 E-07
		GO:0005509	Calcium ion binding	28	8.26	6.10 E-09	7.58 E-07
		GO:0015662	ATPase activity, coupled to transmembrane movement of ions, phosphorylative mechanism	8	1.14	7.50 E-06	6.99 E-04
		GO:0030955	Potassium ion binding	5	0.49	4.80 E-05	2.98 E-03
		GO:0019870	Potassium channel inhibitor activity	5	0.49	4.80 E-05	2.98 E-03
		GO:0019825	Oxygen binding	5	0.65	2.60 E-04	1.39 E-02
		GO:0005267	Potassium channel activity	5	0.76	6.10 E-04	2.78 E-02
		GO:0015077	Monovalent inorganic cation transmembrane transporter activity	9	2.5	6.70 E-04	2.78 E-02
CC	Enriched	GO:0016459	Myosin complex	5	0.46	3.60 E-05	5.35 E-03
		GO:0005882	Intermediate filament	5	0.51	6.90 E-05	5.35 E-03
		GO:0005833	Hemoglobin complex	5	0.56	1.20 E-04	6.20 E-03
		GO:0005856	Cytoskeleton	21	4.61	5.70 E-04	2.21 E-02
		GO:0005578	Proteinaceous extracellular matrix	6	1.18	8.20 E-04	2.54 E-02
		GO:0005834	Heterotrimeric G-protein complex	3	0.26	1.23 E-03	3.18 E-02
Late	Enriched	GO:0006879	Cellular iron ion homeostasis	15	3.05	1.80 E-07	6.89 E-05
		GO:0006826	Iron ion transport	15	3.05	1.80 E-07	6.89 E-05
		GO:0006508	Proteolysis	22	7.2	1.90 E-06	4.85 E-04

MF	Enriched	GO:0006783	Heme biosynthetic process	11	2.1	4.30 E-06	7.05 E-04
		GO:0007602	Phototransduction	5	0.33	4.60 E-06	7.05 E-04
		GO:0018298	Protein-chromophore linkage	5	0.38	1.20 E-05	1.53 E-03
		GO:0007601	Visual perception	10	1	3.00 E-04	3.28 E-02
		GO:0020037	Heme binding	13	2.48	5.80 E-07	2.16 E-04
		GO:0005506	Iron ion binding	16	4.45	6.30 E-06	1.17 E-03
		GO:0009881	Photoreceptor activity	4	0.23	2.30 E-05	2.61 E-03
		GO:0004252	Serine-type endopeptidase activity	8	1.31	2.80 E-05	2.61 E-03
		GO:0004866	Endopeptidase inhibitor activity	14	4.5	1.30 E-04	9.70 E-03
		GO:0004182	Carboxypeptidase A activity	4	0.42	4.90 E-04	3.05 E-02
MF	Depleted	GO:0003746	Translation elongation factor activity	4	0.47	7.90 E-04	3.90 E-02
		GO:0008061	Chitin binding	3	0.23	9.40 E-04	3.90 E-02
		GO:0008533	Astacin activity	3	0.23	9.40 E-04	3.90 E-02
CC	Enriched	GO:0003676	Nucleic acid binding	20	40.18	7.80 E-05	2.91 E-02
		GO:0005576	Extracellular region	15	5.31	2.10 E-04	3.26 E-02

<sup>a</sup>GO ontologies: BP (biological process), MF (molecular function), CC (cellular component).

## 4.9 References

1. Raff RA (1996) The shape of life: genes, development, and the evolution of animal form. Chicago and London: The University of Chicago Press.
2. Galis F, Metz JA (2001) Testing the vulnerability of the phylotypic stage: on modularity and evolutionary conservation. *J Exp Zool* 291: 195-204.
3. Irmeler I, Schmidt K, Starck JM (2004) Developmental variability during early embryonic development of zebra fish, *Danio rerio*. *J Exp Zool B Mol Dev Evol* 302: 446-457.
4. Richardson MK (1995) Heterochrony and the phylotypic period. *Dev Biol* 172: 412-421.
5. Duboule D (1994) Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. *Dev Suppl*: 135-142.
6. Davis JC, Brandman O, Petrov DA (2005) Protein evolution in the context of *Drosophila* development. *J Mol Evol* 60: 774-785.
7. Hazkani-Covo E, Wool D, Graur D (2005) In search of the vertebrate phylotypic stage: a molecular examination of the developmental hourglass model and von Baer's third law. *J Exp Zool B Mol Dev Evol* 304: 150-158.
8. Irie N, Sehara-Fujisawa A (2007) The vertebrate phylotypic stage and an early bilaterian-related stage in mouse embryogenesis defined by genomic information. *BMC Biol* 5: 1.
9. Roux J, Robinson-Rechavi M (2008) Developmental constraints on vertebrate genome evolution. *PLoS Genet* 4: e1000311.
10. Alexeyenko A, Sonnhammer EL (2009) Global networks of functional coupling in eukaryotes from comprehensive data integration. *Genome Res* 19: 1107-1116.
11. Freeman LC (1978/79) Centrality in social networks: conceptual clarification. *Social Networks* 1: 215-239.
12. Pal C, Papp B, Lercher MJ (2006) An integrated view of protein evolution. *Nat Rev Genet* 7: 337-348.
13. Liang H, Li W-H (2009) Functional compensation by duplicated genes in mouse. *Trends in Genetics* 25: 441-442.
14. Sempere LF, Cole CN, McPeck MA, Peterson KJ (2006) The phylogenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint. *J Exp Zool B Mol Dev Evol* 306: 575-588.
15. Thatcher EJ, Flynt AS, Li N, Patton JR, Patton JG (2007) MiRNA expression analysis during normal zebrafish development and following inhibition of the Hedgehog and Notch signaling pathways. *Dev Dyn* 236: 2172-2180.
16. Alexiou P, Maragkakis M, Papadopoulos GL, Reczko M, Hatzigeorgiou AG (2009) Lost in translation: an assessment and perspective for computational microRNA target identification. *Bioinformatics* 25: 3049-3055.
17. Webb SE, Miller AL (2007) Ca<sup>2+</sup> signalling and early embryonic patterning during zebrafish development. *Clin Exp Pharmacol Physiol* 34: 897-904.
18. Whitaker M (2006) Calcium at fertilization and in early development. *Physiol Rev* 86: 25-88.
19. Freisinger CM, Schneider I, Westfall TA, Slusarski DC (2008) Calcium dynamics integrated into signalling pathways that influence vertebrate axial patterning. *Philos Trans R Soc Lond B Biol Sci* 363: 1377-1385.

20. Wienholds E, Kloosterman WP, Miska E, Alvarez-Saavedra E, Berezikov E, et al. (2005) MicroRNA expression in zebrafish embryonic development. *Science* 309: 310-311.
21. Richardson MK (1999) Vertebrate evolution: the developmental origins of adult variation. *Bioessays* 21: 604-613.
22. Solnica-Krezel L (2005) Conserved patterns of cell movements during vertebrate gastrulation. *Curr Biol* 15: R213-228.
23. Cruickshank T, Wade MJ (2008) Microevolutionary support for a developmental hourglass: gene expression patterns shape sequence variation and divergence in *Drosophila*. *Evol Dev* 10: 583-590.
24. Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, et al. (2007) ArrayExpress--a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res* 35: D747-750.
25. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, et al. (2007) Ensembl 2007. *Nucleic Acids Res* 35: D610-617.
26. Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F (2004) A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American Statistical Association* 99: 909-917.
27. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5: R80.
28. Futschik ME, Carlisle B (2005) Noise-robust soft clustering of gene expression time-course data. *J Bioinform Comput Biol* 3: 965-988.
29. Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, et al. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 34: D535-539.
30. Hermjakob H, Montecchi-Palazzi L, Lewington C, Mudali S, Kerrien S, et al. (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res* 32: D452-455.
31. Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, et al. (2006) Human protein reference database--2006 update. *Nucleic Acids Res* 34: D411-414.
32. R Development Core Team (2007) R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.
33. Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, et al. (2004) EnsMart: a generic system for fast and flexible access to biological data. *Genome Res* 14: 160-169.
34. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25: 25-29.
35. Alexa A, Rahnenfuhrer J, Lengauer T (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* 22: 1600-1607.
36. Sprague J, Bayraktaroglu L, Clements D, Conlin T, Fashena D, et al. (2006) The Zebrafish Information Network: the zebrafish model organism database. *Nucleic Acids Res* 34: D581-585.
37. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res* 36: D154-158.
38. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, et al. (2007) NCBI GEO: mining tens of millions of expression profiles--database and tools update. *Nucleic Acids Res* 35: D760-765.

39. Gaidatzis D, van Nimwegen E, Hausser J, Zavolan M (2007) Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics* 8: 69.
40. Bastian F, Parmentier G, Roux J, Moretti S, Laudet V, et al. (2008) Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species. *in* DILS: Data Integration in the Life Sciences. *Lecture Notes in Computer Science*. 5109: 124-131.



# 5 Expression in the nervous system drives retention after whole-genome duplication in vertebrates

---

Julien Roux, Marc Robinson-Rechavi

## Abstract

The evolutionary history of vertebrates is marked by three ancient whole-genome duplications: two successive rounds in the ancestor of vertebrates, and a third one specific to teleost fishes. Biased gene loss of 80-90% of duplicates leads to the enrichment of the genome in certain functions, such as transcription factors, but this selective retention is not fully understood. Especially that there appears to be a complex relation between retention, evolutionary rate, and essentiality.

We used a new method of anatomical ontology enrichment analysis, applied to gene expression data from *in situ* hybridizations of thousands of genes from two vertebrates: zebrafish and mouse. We show that expression in the nervous system drives retention of duplicates after a whole-genome duplication. Further analyses do not seem to support an adaptive explanation for this pattern. Neural structures seem to be simply more tolerant to perturbations such as duplication. We discuss the implications of this result on the previously reported association between rate of sequence evolution and duplicability.

This article is in preparation and should be submitted soon.

## 5.1 Introduction

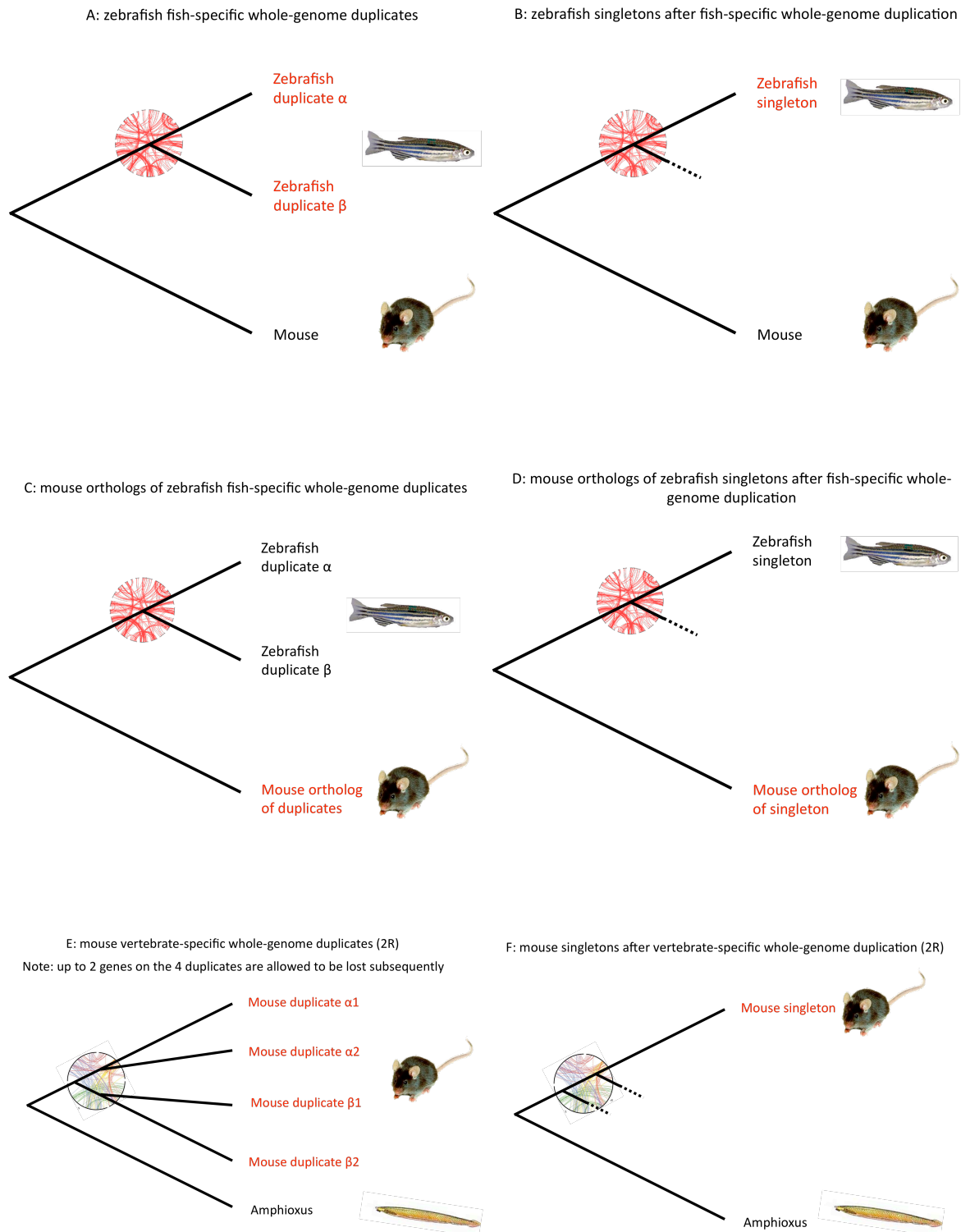
Gene duplication is thought to play a major role in evolution, since it provides raw material for innovation [1]. Whole-genome duplications are special events, doubling all genes of a species at the same time. Such whole-genome duplications occurred quite rarely, but pervasively in the tree of life [2]. Notably, two successive rounds of duplications occurred at the origin of vertebrates [3,4], followed by a third round specific to teleost fishes [5]. The consequences of these events on the evolution of the vertebrate lineage still remain to be identified [2].

The process of gene retention itself is not well understood. After such events, most duplicated genes get rapidly lost. Only 10-20% of the duplicates are retained on the long term [4,6]. These are known to constitute a biased subset of the genome. For instance genes whose sequence evolves slowly appear to be more retained, as well as genes mapped to certain functional categories (e.g. signaling, behaviour, regulation)[4,6,7,8]. Other structural protein features can influence the fate of duplicate genes, such as their length, number of domains, *cis*-regulatory motif or phosphorylation sites [9,10]. Causal relations between these features and the increased propensity of retention after whole-genome duplication have not been yet confirmed.

Recently we have also shown that in vertebrates the pattern of expression through development could also bias duplicate gene retention: genes expressed early in development in zebrafish and mouse tend to be more eliminated [11]. This confirms that patterns of expression strongly influence patterns of molecular evolution in animals. As many studies on whole-genome duplication have been performed on yeast, such biases for vertebrates are somewhat under-studied.

In this paper we use a new bioinformatics method, applied to high quality *in situ* hybridization data, to analyze the retention of duplicate genes regarding expression patterns in anatomical structures. We find that genes expressed in the nervous system have an increased chance of being retained after whole-genome duplication. This pattern is very strong and is observed for the teleost fish specific genome duplication, as well as for the two rounds ancestral to vertebrates ("2R"). It is likely to be explained by a high tolerance of neural tissues to perturbations such as gene duplication. As genes expressed in neural structures are known to evolve slowly [12], a direct prediction of

our results is that the known correlation between protein evolutionary rate and duplicate retention could be spurious. In fact, we show that expression in the nervous system, duplicability, and evolutionary rates interact in a complex manner.



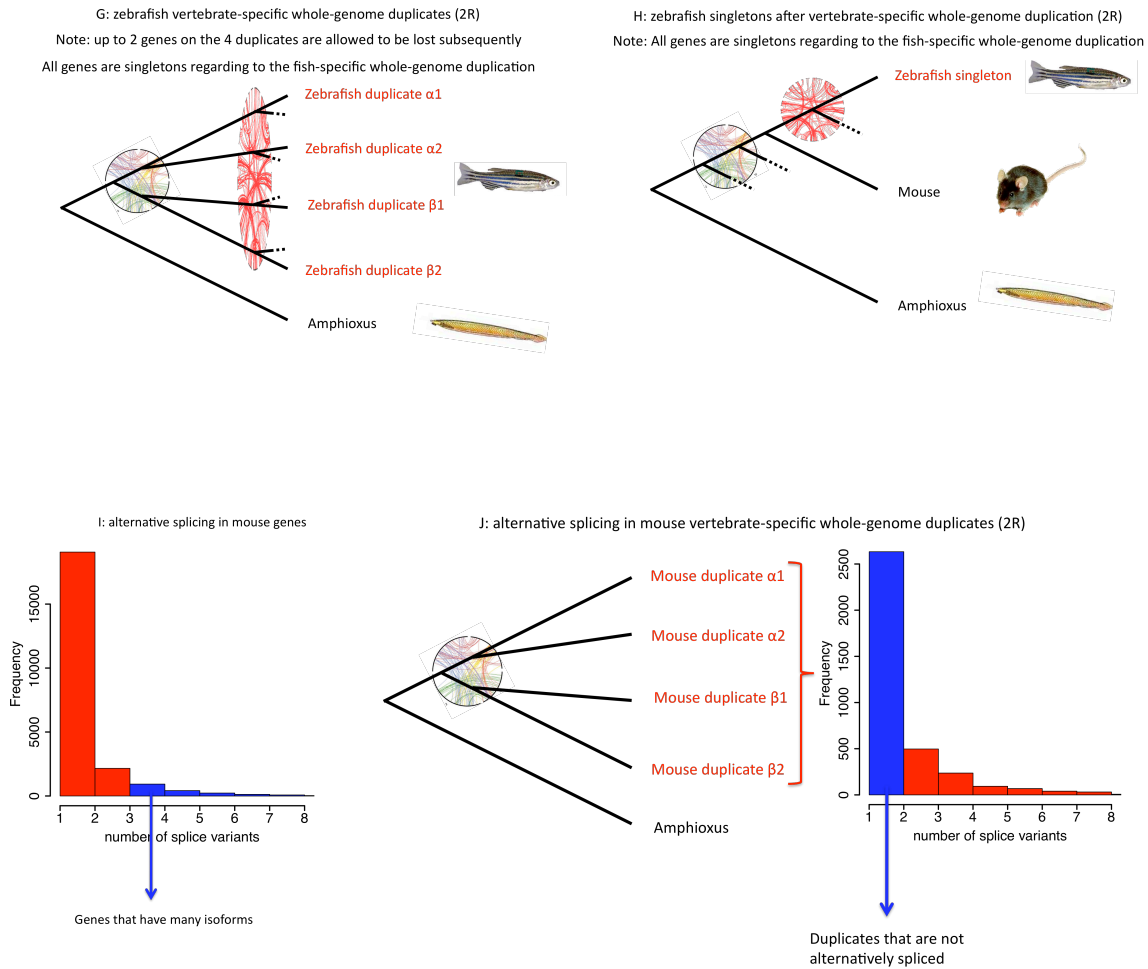


Figure 1: explanatory figure for the different evolutionary scenarios experienced by the groups of genes tested in the article

## 5.2 Results

### *Fish-specific whole-genome duplication*

To analyze potential anatomical retention bias of duplicates after the fish-specific whole-genome duplication (Figure 1A; Table 1[i]), we collected zebrafish *in-situ* hybridization expression data. This technique allows a precise mapping of expression to fine anatomical structures, that microarray studies cannot perform and thus tend to average [32]. Several thousands of *in-situ* hybridizations have been performed in zebrafish, allowing us to use these data to analyze trends at the genomic scale. These expression data are mapped to an ontology describing the anatomy of zebrafish [18]. Similarly to the widely used Gene Ontology enrichment test [33,34], we performed a Fisher test for each category of the zebrafish anatomical ontology. This test compares the proportion of duplicate genes mapped to a given category (i.e. expressed in a given

tissue) to the proportion in the reference set (“universe”). Here, the universe is composed of all genes having *in-situ* expression data.

The list of anatomical structures showing an enrichment of expression of duplicates from the fish-specific whole-genome duplication is shown in Table S1. With a false discovery rate (FDR) threshold of 20%, 224 structures are significant. This means that more duplicate genes are expressed in these structures than expected by chance. Many of these are high-level terms in the ontology (e.g. “anatomical system” or “organism subdivision”), probably due to correlation between categories: high-level structures benefit from the propagation of the expression mapped to all their sub-structures in the ontology. This problem has been acknowledged concerning Gene Ontology tests [34] and several decorrelation algorithms have been developed to reduce this type of local dependencies [24,35,36]. We used here the “elim” algorithm available in the package topGO, a bottom-up approach that stops the propagation of genes mapped to significant categories to higher level terms [24]. This test yields a less redundant list of 117 anatomical terms where duplicates are more expressed than expected (Figure 1A; Table 1[ii]). It is shown in Table 2. Of note, similar results were obtained using the database Homolens 3 [21] to detect duplicate genes (not shown; see Methods).

It is clear that these 117 terms include many structures from the nervous system. For instance the top five enriched structures are “retinal ganglion cell layer”, “spinal cord”, “neuron”, “epiphysis” and “tegmentum” (Table 2). Because a high number of genes are known to be expressed in the nervous system [32], we tested if the high proportion seen in Table 2 is significant. We built a reference dataset gathering all nervous system structures. As it is sometimes difficult to classify a structure as nervous or not, we created two datasets, a “broad” one, including notably sensory systems as well as embryonic precursors of nervous structures, and a more strictly defined “narrow” dataset (see Methods). Using both datasets, we found a highly significant enrichment of nervous system structures in Table 2 (Fisher tests; “broad”:  $p=5.3e-11$ , with odds-ratio=3.6; “narrow”,  $p=7e-05$  with odds-ratio=2.3). Changing the threshold for reporting anatomical terms did not alter the results (e.g. using a stringent FDR of 5%; “broad”:  $p=1.1e-6$ , odds=4.6; “narrow”:  $p=0.005$ , odds=2.5). The results with no decorrelation

algorithm (Table S1) also yield a significantly enriched proportion of nervous system structures (“broad”:  $p=3.6e-15$ , odds=3.1; “narrow”:  $p=8.3e-7$ , odds=2.2).

We observe that the use of the “broad” list of nervous system structures gives lower p-values than with the “narrow” list, suggesting that duplicates are expressed quite broadly in the nervous system (including in developing precursors and sensory organs). Indeed the significant terms represent very diverse nervous system structures and no specific bias inside the nervous system is apparent.

Because some of the significant structures do not belong to the nervous system (e.g. heart, FDR = 0.0069), we applied the same procedure to other anatomical systems to test if they were also over-represented in Table 2. Eight systems are easily isolated using the structure of the zebrafish anatomical ontology: respiratory system, cardiovascular system, renal system, digestive system, skeletal system, musculature system, reproductive system and immune system. But no significant over-representation of these systems was found on the list, most of them being depleted (e.g. skeletal system,  $p=3e-5$ , odds=0.15).

We next performed the same analysis on singleton genes (genes whose duplicates were lost after the fish-specific whole genome duplication; Figure 1B; Table 1[iii]). This yielded only 3 structures enriched in expression of this group: “alar plate midbrain”, “unspecified” and “liver”.

However, we found 82 structures depleted in expression of singletons (Table S2), most of them part of the nervous system (Fisher test; “broad” list:  $p=9.5e-7$ , odds=3.1). Interestingly using the “narrow” list did not yield a quite significant p-value, confirming that developmental precursors of the nervous system and peripheral structures such as sensory organs play an important role in the trend (Fisher test;  $p=0.1$ , odds=1.5). For clarity, in the rest of the article we report only values obtained using the “broad” dataset, unless mentioned explicitly.

To summarize, we observe that the genes retained in duplicate after the fish-specific whole-genome duplication are more expressed in nervous system structures than expected by chance, whereas the genes whose duplicates were not retained (singletons) are less expressed in these structures than expected.

### ***Bias in retention or evolution after duplication?***

Two scenarios can explain this result. First, retention of two copies may be more likely after the whole genome duplication for genes less expressed in nervous system structures. Second, the retention of genes may be unbiased relative to expression, but duplicate genes may evolve secondarily expression in the nervous system. To get a proxy of the ancestral state before whole-genome duplication, we used mouse data, which has diverged from zebrafish before the fish specific duplication. Since a large number of *in-situ* hybridization data are also available for mouse (see Methods), we could apply strictly the same methodology as for zebrafish to detect biases in expression patterns.

We compared mouse orthologs of zebrafish duplicates, to mouse orthologs of zebrafish singletons, regarding their expression pattern (Figures 1C and D; Table 1[iv] and [v]). Mouse orthologs of duplicates were found more expressed than expected in nervous system structures ( $p=0.0001$ , odds=6). This result in mouse is consistent with the observations in zebrafish, and the most parsimonious explanation is that expression was similar in the ancestor of the two lineages. Therefore the first scenario is most probable: after the fish specific whole genome duplication, there was preferential retention of duplicates expressed in the nervous system.

Interestingly the mouse orthologs of singletons show no depletion of expression in nervous structures, as was observed for zebrafish singletons. This might be due to zebrafish singletons having evolved secondarily a lower expression in the nervous system after the whole-genome duplication event.

### ***Vertebrate whole-genome duplications***

To check if this bias is specific to the fish-specific genome duplication, we repeated the analysis with the two ancient rounds of genome duplication ("2R"), which occurred in the ancestor of vertebrates [4]. It is difficult to distinguish between the two whole genome duplications since no model species diverged from the vertebrate lineage between them. Therefore we looked at the genes with any duplication at the origin of vertebrates (Figure 1E; Table 1[vi]). The pattern uncovered is very similar to that of fish-specific duplicates. In mouse, most of the structures enriched in expression of vertebrate-specific duplicates are part of the nervous system ( $p=2.5e-6$ , odds=4.7). As observed in zebrafish, this is the case also for structures depleted in expression of

singletons (Figure 1F; Table 1[vii]), although the p-value is not quite significant ( $p=0.077$ , odds=2.5).

Of note, these results are consistent if we use an independent source of *in situ* hybridization data: the project Eurexpress performs a systematic screening of gene expression in wild-type mouse embryos at E14.5 (see Methods). The annotation is made coherently by one research group of the consortium. The enrichment test on 2R whole-genome duplicates yields fewer results (10 significant terms), because less data are available, but similar results are obtained (the 10 significant terms are substructures of the nervous system; not shown).

Zebrafish also experienced this duplication (Figures 1G and H; Table 1[viii] and [ix]), but only 2 structures are found enriched in 2R duplicates (“angioblastic mesenchymal cell” and “spinal cord neural rod”), while 2R singletons show a significant depletion of expression in nervous system structures ( $p=0.0028$ , odds=1.6). This might be due to our use only of genes which duplicated in 2R, but not in the fish specific genome duplication. Thus it seems that the mechanisms of retention after whole genome duplication are conserved during vertebrate evolution.

### ***Recent species-specific duplications***

Duplicate genes can arise from other sources than whole genome duplications. In this case, bias of retention is acting jointly with other mechanisms, such as bias of generation or of fixation of duplicates.

Concerning lineage-specific single gene duplicates in zebrafish and mouse (mostly recent tandem duplications), we do not observe any enrichment of expression in nervous system structures (Table 1[x] and [xi]). In mouse however we detect a significant depletion of expression in the nervous system of these recent duplicates ( $p=0.0032$ , odds=9.6). This seems to be opposite to the trend for whole-genome duplications, and may reflect the fact that small-scale duplications and whole-genome duplications affect different genes [31,37].

### ***Number of isoforms***

Yeo et al [38] have shown that an unusually high frequency of conserved human-mouse alternative splicing is present in genes expressed in the brain. This led us to test if other



modes of sequence evolution than gene duplication could also be favoured for genes expressed in the nervous system. In mouse we indeed detect that genes with at least 3 isoforms (Figure 1I; Table 1[xii]) are enriched in nervous system expression ( $p=0.00019$ , odds=  $\infty$  since all the 6 significant structures are from the nervous system). Results in zebrafish are in the same direction but are not significant, probably because of a lower EST coverage and thus less recognized splice variants in this species ( $p=0.17$ , odds=2.5).

This result is surprising since alternative splicing and gene duplication have been reported to be anti-correlated mechanisms [39,40]. More detailed analysis shows that the anti-correlation does not hold for old genes, such as those that experienced the 2R whole-genome duplications (Roux and Robinson-Rechavi, unpublished); these genes show a positive correlation between alternative splicing and duplication. This might explain why genes with many isoforms are enriched in nervous system expression.

Indeed we confirmed that high number of alternative splice variants is not a causal factor by keeping only genes that were retained in duplicate after 2R, but have no alternative splicing (1 isoform; Figure 1J; Table 1[xiii]). These genes also show a marginally significant enrichment of expression in nervous structures ( $p=0.06$ , odds=4.3).

### ***Nervous system expression and rate of sequence evolution***

The sequence of genes expressed in neural tissues tends to evolve slowly [12,41,42]. This is hypothesized to be due to a high sensitivity of neurons to protein synthesis errors resulting in protein misfolding. Misfolded proteins can be toxic to cells because they are prone to aggregate with other misfolded proteins and to hydrophobic surfaces like membranes. Because of their long lifetimes and high membrane surface-area, selection to prevent misfolding is very strong in neurons. Amino acid changes are thus prevented since they are likely to increase the propensity of a protein to misfold, and genes expressed in neural tissues consequently display a slow rate of sequence evolution.

To check this pattern with our dataset, we isolated the 10% slowest evolving genes in mouse and looked in which tissues their expression was enriched (Table 1[xiv]). Rates of evolution were measured using the  $d_N/d_S$  ratio. As expected, among the structures

expressing preferentially slowly evolving genes, we find a very strong enrichment of nervous system structures ( $p=1.4e-7$ , odds=4.5). Among other anatomical systems, the skeleton is marginally enriched ( $p=0.088$ , odds=2).

Selection against protein misfolding was also reported to act at synonymous sites, biasing codon usage [12,27]. To improve translation accuracy and prevent misfolding, the structurally important sites of a protein are often observed to be encoded by a codon matching perfectly its cognate tRNA.

To check this trend with our dataset, we looked at two different measures. First, expression of the 10% of genes with lowest  $d_s$  (Table 1[xv]) is marginally enriched in the nervous system ( $p=0.073$ , odds=1.9). Interestingly, a significant enrichment is found using the “narrow” dataset ( $p=0.0092$ , odds=2.8), as well as in the renal and urinary system ( $p=0.014$ , odds=6.6) and especially in the skeletal system ( $p=1.8e-7$ , odds=7.5). Many mesenchymal cartilage condensations show an enrichment of expression of genes with low  $d_s$ . This has never, to our knowledge, been previously reported. As these structures are patterned by the action of morphogens[43], it might be possible that such genes are selected for high efficiency of translation, constraining strongly their synonymous sites. Further investigations are required to understand this pattern.

Secondly, we calculated Akashi’s score of optimal codon use (Psi value) for all mouse genes (see Methods), and looked at anatomical structures enriched in expression of the 10% of genes showing the strongest effect (Table 1[xvi]). “Broad” nervous system is now strongly enriched ( $p=3.6e-10$ , odds=4.7), while the “narrow” dataset is not significant ( $p=0.27$ , odds=1.4). Renal and urinary system is again enriched ( $p=0.019$ , odds=4.3), but skeleton is this time under-represented ( $p=0.061$ , odds=0.28).

This pattern, although complex, seems to indicate that the expression of genes in specific anatomical structures, including but not exclusively neural tissues, can lead to selection on synonymous mutations. This is important to underline as such selection is thought to be weak in mammalian genomes [42,44].

### ***Explaining duplicate retention bias***

Several hypotheses to explain gene duplicate retention involve selection for increased gene dosage. In yeast, this effect has been shown to be not significant [45]. In mammals,

some evidence suggests that many genes are expressed in the brain, but at a rather low level [32], in contradiction with the gene dosage hypothesis.

But other constraints on optimal expression level have been shown to influence gene duplication retention. This is for example the case of genes belonging to metabolic pathways [46], or for genes expressed very early in development [11].

To test such constraints on dosage, we looked at potential constraints on gene loss in specific tissues. We isolated the set of essential genes in mouse, whose knock-out is lethal or leads to sterility. We looked at the anatomical structures enriched in the expression of essential genes. The “universe” we took as reference is the set of genes with expression data and reported knock-out phenotype (1923 genes; Table 1[xvii]).

Essential genes, when compared to this appropriate reference, are not found significantly enriched in many early embryonic precursor tissues (e.g. mesoderm, mesenchyme, somite, endoderm). But no particular anatomical system is more represented among them (e.g. nervous system “broad”:  $p=0.4$ , odds=1.4). This supports the idea that the main factor influencing essentiality in vertebrates is early timing of expression during development [11,31]. Genes expressed later in development, including those expressed in the nervous system, do not seem particularly constrained concerning gene loss of function.

### ***Are duplicates slowly evolving genes?***

The relationship between gene duplication and expression in the nervous system questions previous observations, that genes kept in duplicate after whole-genome duplications are a slowly evolving subset [6,7]. This relation might be spurious, a consequence of slow sequence evolution of genes expressed in the nervous system [12,41].

We first confirm with our dataset that mouse genes kept in duplicate after the vertebrate-specific genome duplication have a lower  $d_N/d_S$  than singletons (Wilcoxon test,  $p < 2.2e-16$ ). Similarly we confirm that genes expressed in the nervous system have a lower  $d_N/d_S$  than genes which are not expressed in the nervous system (Wilcoxon test,  $p=3e-11$ ).

We then split our dataset into 4 categories according to the two levels of the two factors (duplicates expressed in the nervous system or not, and singletons expressed in the nervous system or not). First, the overall variation of rate of sequence evolution observed in the 4 groups is significant (Kruskal-Wallis test,  $p < 2.2e-16$ ; Figure 2). Second, a significant difference between duplicates and singletons is found for genes expressed in the nervous system (Wilcoxon test,  $p = 0.00023$ ), but not for genes that have no expression in the nervous system ( $p = 0.85$ ). This indicates that there is a relation between duplication and the rate of sequence evolution, but only for genes expressed in the nervous system.

Third, the expected slower evolution of nervous system genes compared to non nervous system genes is valid for duplicates ( $p = 0.00012$ ), but not for singletons ( $p = 0.69$ ). This is surprising and it seems difficult to explain why selection against protein misfolding would not apply to singletons.

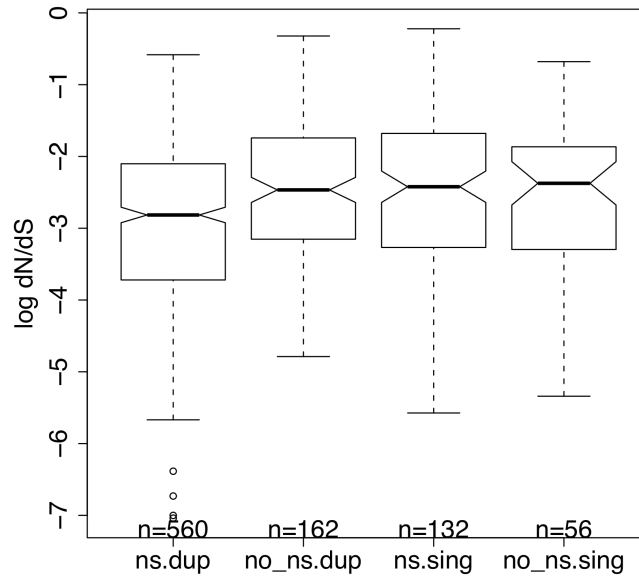


Figure 2: comparison of the rate of protein sequence evolution ( $\log d_N/d_S$ ) for genes kept in duplicate (“dup”) or not (singletons, “sing”) after the vertebrate whole-genome duplications, depending on their expression on the nervous system (“ns” and “no\_ns”).

Globally this analysis shows that both the relation between duplicability and rate of evolution, and between expression in the nervous system and rate of evolution, are partially spurious, and that an interaction between duplicability and expression in the nervous system should be considered in this context.

### 5.3 Discussion

In this study we have taken advantage of high quality *in situ* hybridization data allowing precise description of gene expression patterns in mouse and zebrafish. These are mapped on ontologies describing precisely the anatomy of these species, making it possible to perform ontology enrichment tests and isolate the tissues showing enrichment in expression of genes of interest. This methodology corrects for potential annotation biases and biases of the subset of genes for which expression data are available [34].

We uncover a strong trend for genes to be retained after whole-genome duplication if they are expressed in neural tissues. This pattern was not uncovered previously, probably because of the high complexity of gene expression in the nervous system [32].

This is however consistent with reports of Gene Ontology annotations significantly enriched in whole-genome duplicates [4,6,11].

The high rate of duplication in neural tissues stands in contrast to observations concerning non-synonymous and synonymous mutations. Genes expressed in the nervous system experience strong constraints preventing changes in their sequence [41]. This has been explained by selection against protein misfolding, which is felt strongly on these non-regenerating tissues [12]. The contagious aggregation of misfolded proteins is found to be the reason for many neurodegenerative disorders [47].

Classical models of duplicate gene retention are sub-functionalization [48] and neo-functionalization [49]. Duplicate genes are kept because they share the ancestral function and become indispensable, or if one gene diverges and acquire a new function. Some more complex intermediate cases have been also proposed [50]. It is possible that expression in the nervous system can make these processes easier. The high number of cell types present in the nervous system could possibly play such a role [51]. Both sub- and neo-functionalization imply first an accelerated rate of sequence evolution, due to relaxation of purifying selection on both duplicates, or to positive selection on one of the duplicates. But the strong constraints on sequence for genes expressed in the nervous system seem to make this less likely. Duplication probably cannot reduce this burden since it will increase the dosage of the protein product, increasing even more selection against protein misfolding.

Divergence of expression pattern (due to modifications in *cis*-regulatory regions) might also help sub- and neo-functionalization, but it has been shown that the divergence between duplicates expressed in the nervous system is very low [41]. Similarly most highly conserved enhancers seem to drive expression in the nervous system ([52]; Slavica and Bucher, personal communication).

Another hypothesis for higher duplicability of genes expressed in neural tissues could be selection for an optimal level of expression [11,46]. In this case a correlation should be seen between the propensity of duplication and the gravity of phenotype when genes are lost (after knock-out for example). Both gene duplication and gene loss can be seen as a perturbation for the organism [11,53]. This is not what we observe for genes

expressed in the nervous system, which do not seem to induce significantly more serious phenotypes than other genes after knock-out.

A non-adaptive hypothesis thus seems to be more likely: neural tissues might simply be more tolerant than other tissues to changes of gene dosage by gene duplication. If duplication is viewed as a perturbation, the genes retained in duplicates will be those imposing the smallest perturbation to the organism [11,54]. Nervous system may be more robust, and accept more easily such perturbations. Coherently with this hypothesis, a continuous addition of transposable elements to the somatic genome of neural cells has been reported; such perturbations are tolerated without apparent phenotypic change [55].

The implications of our results are manifold. First, the observation that genes kept in duplicate after a whole-genome duplication have a slow rate of sequence evolution [6,7] is partially explained by the slow rate of evolution of nervous system genes, which are preferentially retained in duplicate.

Second, we show that essential proteins are not over-expressed in neural tissues and thus are unlikely to experience strong selection against protein misfolding. This may contribute to the weakness of the correlation between essentiality and rate of sequence evolution [56]. The small but significant effect detected might be due to the functional constraints related to development, as essential genes are expressed early during development in vertebrates, where some level of constraint on protein sequence has been observed [11].

Third, as gene retention is high in the nervous system, we predict that many pairs of gene duplicates will show expression of one duplicate in the nervous system and one duplicate somewhere else. The interpretations of such patterns should be made with caution, particularly the scenarios involving adaptive specialization after duplication.

Finally, if the increased retention of duplicate genes expressed in the nervous system is not adaptive, it should not be taken as a direct explanation for the complexity of the vertebrate nervous system. It is possible however that genome duplications contributed to expand the toolkit of genes available in the nervous system, that could be co-opted later to evolve new vertebrate-specific features, such as the neural crest [57].

## 5.4 Material and Methods

### ***Mouse expression data***

*Mus musculus* RNA *in-situ* hybridization expression data were retrieved from the GXD database on June 2009 [13]. Wild-type data, obtained under non pathological conditions, with no treatment (“normal” gene expression) were integrated into Bgee (<http://bgee.unil.ch/>), a database allowing the comparison of transcriptome data between species [14]. Expression data are mapped to anatomical ontologies. Bgee uses an abstract version of the mouse embryo anatomical ontology (EMAPA)[15], adapted from the EMAP ontology [16]. A correspondence file between both ontologies can be found at [http://bgee.unil.ch/bgee/download/mapping\\_EMAP\\_to\\_EMAPA.txt](http://bgee.unil.ch/bgee/download/mapping_EMAP_to_EMAPA.txt). The data used in this article are from the release 6 of Bgee (September 2009).

An independent dataset was retrieved from Eurexpress (<http://www.eurexpress.org/>), a consortium creating expression data of more than 20,000 mouse genes by RNA *in situ* hybridization on sagittal sections from E14.5 wild-type embryos. The annotation of the data and the mapping on the mouse embryo anatomical ontology is made coherently by only one lab. Data were retrieved using Biomart [17] on the Eurexpress website.

### ***Zebrafish expression data***

*Danio rerio in-situ* hybridization expression data were retrieved from the ZFIN database on June 2009 [18]. Wild-type data, obtained under non pathological conditions, with no treatment (“normal” gene expression) were integrated into Bgee. We used the zebrafish anatomical and developmental ontology developed by ZFIN [18]. The data used in this article are from the release 6 of Bgee (September 2009).

### ***Identification of duplicate genes***

Gene families were obtained from the Ensembl database release 56 [19]. We used the Perl API to query the Ensembl Compara Gene trees [20] and scan for specific gene topologies. We selected sets of genes with or without duplications on specific branches of the vertebrate phylogenetic tree. The perl scripts used are available upon demand.

Regarding the fish-specific whole genome duplication in zebrafish, we found 3769 Ensembl gene IDs for duplicates, 8995 for singletons, 1732 mouse orthologs of fish duplicates, and 9011 mouse orthologs of fish singletons. For the 2R whole genome



duplications, we found 1210 duplicates and 2867 singletons in zebrafish, and in mouse 3629 duplicates, 1907 singletons with no later duplications and 2812 singletons with later duplications. We also retrieved 5494 recent lineage-specific duplicates in zebrafish and 2378 in mouse. Recent duplicates with 0 or 100% sequence identity were removed from the dataset, because they probably correspond to assembly artefacts.

An independent analysis was performed using gene families obtained from the HomolEns database version 4 (<http://pbil.univ-lyon1.fr/databases/homolens.html>) [21], using the TreePattern functionality of the FamFetch client for HomolEns, which allows scanning for gene tree topologies [22].

### ***Ontology enrichment analyses***

Enrichment and depletion of expression in anatomical were tested with a Fisher exact test using a modified version of the R and Bioconductor package topGO (<http://www.R-project.org/>; <http://bioconductor.org/>) [23,24,25] (Adrian Alexa, personal communication), allowing to handle ontologies in OBO format (<http://www.obofoundry.org>) [26]. We defined the reference set as all the genes for which we had expression data in at least one structure of the organism during development. This accounted for 7957 genes in zebrafish and 4915 genes in mouse in respectively 966 and 1510 different anatomical structures. The “elim” algorithm was used because it allows to decorrelate the ontology graph structure (parent-child relationships), reducing non-independence problems of classical tests. In the algorithm a Fisher test is performed on the contingency tables. A False Discovery Rate correction was applied; ontology categories enriched or depleted with a FDR < 20% are reported.

### ***List of nervous system anatomical structures***

Two reference lists of nervous system organs were extracted from the ontologies for zebrafish and mouse. Because it was sometimes hard to choose objectively if a structure belongs to nervous system or not (e.g. sensory organs), we created a “broad” list and a “narrow” list. In zebrafish, the “narrow list” includes the following structures, as well as their sub-structures in the ontology: “nervous system” (ZFA:0000396), “nerve” (ZFA:0007009), “neuroendocrine cell” (ZFA:0009098) and “neuroepithelial cell” (ZFA:0009306). “Sensory system” organs (ZFA:0000282) were removed. The “broad” list includes them, as well as presumptive neural structures during development and

their sub-structures (ZFA:0000063: “presumptive neural plate”, ZFA:0000132: “neural plate”, ZFA:0000133: “neural rod”, ZFA:0000131: “neural keel”, ZFA:0001135: “neural tube”, ZFA:0000045: “neural crest”, ZFA:0001120: “neuroectoderm”, ZFA:0001071: “presumptive neural retina”, ZFA:0001334: “presumptive enteric nervous system”, ZFA:0009012: “neuroplacodal cell”, ZFA:0009080: “neurectodermal cell”, ZFA:0009150: “Rohon-Beard neuron”, ZFA:0001082: “chordo neural hinge”).

In mouse, the “narrow list” includes the following structures, as well as their sub-structures in the ontology: “nervous system” (EMAPA:16469) and “tail nervous system” (EMAPA:16753). We removed “future brain” (EMAPA:16471), “future spinal cord” (EMAPA:16525) and “future spinal cord” (under “tail”; EMAPA:16755) and their sub-structures from this dataset, but included them in the “broad” list, as well as “sensory organs” (EMAPA:16192), “neural ectoderm” (EMAPA:16073) and their sub-structures.

### ***List of anatomical structures from other systems***

We selected the high-level terms in the ontologies corresponding to broad anatomical systems. Because of different structures of the ontologies of zebrafish and mouse ontologies, we could not select exactly the same systems (for example immune system is not present in the mouse ontology). We then retrieved all the terms under these high-level terms down to the leaves of the ontology. In zebrafish we retrieved all organs corresponding to the following systems: respiratory system (ZFA:0000272), cardiovascular system (ZFA:0000010), renal system (ZFA:0000163), digestive system (ZFA:0000339), skeletal system (ZFA:0000434), musculature system (ZFA:0000548), reproductive system (ZFA:0000632) and immune system (ZFA:0001159).

In mouse, we retrieved all organs corresponding to the following systems: skeleton (EMAPA:17213), cardiovascular system (EMAPA:16104), integumental system (EMAPA:17524), alimentary system (EMAPA:16246), respiratory system (EMAPA:16727), renal/urinary system (EMAPA:17366), reproductive system (EMAPA:17381) and liver and biliary system (EMAPA:16840). No high-level term gathers muscular system organs, so we chose one of the biggest node in the ontology linked to muscle: vertebral axis muscle system (EMAPA:17743).

### ***Number of isoforms***

We retrieved the number of different transcripts for mouse protein coding genes from Ensembl 56 [19], using BioMart [17] (attribute “transcript\_count”).

### ***Rate of sequence evolution***

We retrieved the  $d_N$  and  $d_S$  measures for mouse genes (using one-to-one orthologs in rat), from Ensembl 56 [19], using BioMart [17].

### ***Akashi's test***

Selection for translational accuracy was tested using Akashi's test [12,27]. Alignments of mouse and rat protein-coding genes were retrieved from Ensembl using the Perl API. Sites with the same amino acid at the aligned position in mouse and rat orthologous gene sequences were designated conserved. Optimal codons in mouse were taken from Drummond and Wilke [12]. Laplace smoothing (or estimate) was applied to contingency tables in order to remove problems with counts of zero. The outputs of the test are: (i) a Z-score, which assesses how likely the association in a gene sequence between conserved sites and preferred codons is to have occurred by chance (significance); we assume that the global Z-score for a group of genes follows the standard normal distribution, so that a p-value can be computed (e.g.  $p(Z \geq 1.96) = 0.025$ ); (ii) a Psi-score that assesses how strong is the association between preferred codons and conserved sites, which is computed as an odds ratio.

### ***Mouse phenotypes***

Data on mouse mutants were retrieved from the Mouse Genome Database (<ftp://ftp.informatics.jax.org/pub/reports/index.html>, Mars 2010) [28]. We extracted from the files MGI\_Phen GenoMP.rpt and MGI\_PhenotypicAllele.rpt all informations on genotypes and their phenotype for alleles mapped to Ensembl genes. As in Liao and Zhang [29], Liang and Li [30], Makino et al. [31], we called essential those genes giving a lethal phenotype or sterility (upper phenotype categories MP:0005374, MP:0005373, MP:0001924, MP:0001730 and MP:0002083 and their children). See [http://www.informatics.jax.org/searches/MP\\_form.shtml](http://www.informatics.jax.org/searches/MP_form.shtml) for information on phenotypic categories.

We filtered on the technique used and kept only the single mutants obtained with a targeted knock-out. We obtained 2063 essential genes, and 1102 of them had expression data.

## **5.5 Acknowledgements**

We thank members of the Robinson-Rechavi lab, Alain Prochiantz and Allan Drummond for helpful discussions, Adrian Alexa for help in the adaptation of the topGO package, the ZFIN team for providing *in situ* hybridization data and modifications of download files on their website, and the GXD team for providing *in situ* hybridization data and help for the database connection. We acknowledge funding from Etat de Vaud and Swiss National Science Foundation grant 116798.

## **5.6 Tables**

*Table 1: Overview of the different tests performed in the article, and links to explanatory figures. In italics are shown marginally significantly enriched anatomical systems.*

Species	List of genes tested	Significant structures enriched	Anatomical systems enriched	Significant structures depleted	Anatomical depleted	systems	Explanatory figure	List shown in table
[i] zebrafish	duplicates after fish-specific whole-genome duplication	117	Nervous system "narrow"	0	-		1A	S1
[ii] zebrafish	duplicates after fish-specific whole-genome duplication [no decorrelation]	224	Nervous system "broad"; nervous system "narrow"	0	-		1A	2
[iii] zebrafish	singletons after fish-specific whole-genome duplication	3	-	82	Nervous system "broad"; nervous system "narrow"		1B	S2
[iv] mouse	orthologs of zebrafish duplicates after fish-specific whole-genome duplication	20	Nervous system "broad"; nervous system "narrow"	0	-		1C	
[v] mouse	orthologs of zebrafish singletons after fish-specific whole-genome duplication	13	-	0	-		1D	
[vi] mouse	duplicates after vertebrate-specific whole-genome duplication	39	Nervous system "broad"; nervous system "narrow"	0	-		1E	
[vii] mouse	singletons after vertebrate-specific whole-genome duplication	0	-	16	Nervous system "broad"		1F	
[viii] zebrafish	duplicates after vertebrate-specific whole-genome duplication	2	-	0			1G	
[ix] zebrafish	singletons after vertebrate-specific whole-genome duplication	5	-	168	Nervous system "broad"; Nervous system "broad"; Nervous system "narrow"		1H	
[x] mouse	Recent species-specific duplicates	0	-	8	nervous system "narrow"			
[xi] zebrafish	Recent species-specific duplicates	1	-	1	-			
[xii] mouse	genes with at least 3 isoforms	6	Nervous system "broad"; nervous system "narrow"	0	-		1I	
[xiii] mouse	duplicates after vertebrate-specific whole-genome duplication with at least 3 isoforms	7	Nervous system "broad"; nervous system "narrow"	0	-		1J	
[xiv] mouse	10% lowest $d_N/d_S$	52	Nervous system "broad"; nervous system "narrow"; skeletal system Nervous system "broad"; nervous system "narrow"; renal and urinary system;	0	-			
[xv] mouse	10% lowest $d_S$	35	skeletal system	0	-			
[xvi] mouse	10% highest Akashi's test scores	70	Nervous system "broad"; renal and urinary system	0	-			
[xvii] mouse	essential genes	30	embryonic structures (*)	0	-			

(\*) No statistical test performed because structures are not gathered under a common node in the mouse anatomical ontology.

Table 2: List of anatomical structure showing a significant enrichment in expression of genes kept in duplicate after the fish-specific genome duplication (3R; FDR < 20%). In bold the sub-structures of the nervous system (“broad”), and in italics the sub-structures of the nervous system with sensory organs and developmental precursors (“narrow”).

Organ ID	Organ name	Total genes expressed	Duplicates expressed	Duplicates expected	p-value	FDR
ZFA:0000024	<b>retinal ganglion cell layer</b>	307	107	50.35	6.86E-16	7.65E-13
ZFA:0000075	<i>spinal cord</i>	860	225	141.05	6.83E-15	3.81E-12
ZFA:0009248	<i>neuron</i>	562	168	92.17	7.40E-13	2.75E-10
ZFA:0000019	<b>epiphysis</b>	400	119	65.6	5.68E-12	1.58E-09
ZFA:0000160	<i>tegumentum</i>	435	125	71.34	2.20E-11	4.90E-09
ZFA:0000029	<i>hindbrain</i>	1189	289	195	4.67E-11	8.68E-09
ZFA:0000013	<i>cranial ganglion</i>	487	159	79.87	1.57E-10	2.49E-08
ZFA:0000119	<b>retinal inner nuclear layer</b>	213	72	34.93	2.30E-10	3.20E-08
ZFA:0000402	<i>olfactory bulb</i>	79	35	12.96	4.28E-09	5.30E-07
ZFA:0000079	<i>telencephalon</i>	730	190	119.72	1.39E-08	1.55E-06
ZFA:0000761	<b>basal plate midbrain</b>	250	72	41	4.24E-07	4.30E-05
ZFA:0000162	<b>trigeminal placode</b>	177	55	29.03	7.44E-07	6.91E-05
ZFA:0000295	<i>trigeminal ganglion</i>	110	39	18.04	8.16E-07	7.00E-05
ZFA:0000155	somite	969	212	158.92	1.22E-06	9.68E-05
ZFA:0000101	<i>diencephalon</i>	1140	294	186.97	1.91E-06	0.000137467
ZFA:0007007	<i>pallium</i>	26	15	4.26	1.97E-06	0.000137467
ZFA:0000047	<b>peripheral olfactory organ</b>	462	116	75.77	4.43E-06	0.000290815
ZFA:0000149	primitive heart tube	97	34	15.91	5.48E-06	0.000339628
ZFA:0000152	<b>retina</b>	1278	323	209.6	7.94E-06	0.000465778
ZFA:0009150	<b>Rohon-Beard neuron</b>	46	20	7.54	1.29E-05	0.000719957
ZFA:0001056	myotome	622	141	102.01	1.46E-05	0.000776859
ZFA:0000105	epidermis	327	82	53.63	2.74E-05	0.00137367
ZFA:0000143	<b>retinal photoreceptor layer</b>	108	35	17.71	2.83E-05	0.00137367
ZFA:0000120	<b>lateral line ganglion</b>	115	44	18.86	3.16E-05	0.001469892
ZFA:0001314	<b>posterior lateral line ganglion</b>	31	15	5.08	3.39E-05	0.001510146
ZFA:0000051	otic vesicle	649	144	106.44	3.77E-05	0.001616723
ZFA:0000003	adaxial cell	316	79	51.83	4.30E-05	0.001732689
ZFA:0000778	<i>interneuron spinal cord</i>	25	13	4.1	4.35E-05	0.001732689
ZFA:0000048	<b>olfactory placode</b>	364	88	59.7	6.06E-05	0.002329404
ZFA:0009053	<b>sensory neuron</b>	5	5	0.82	0.000117902	0.004382012
ZFA:0000028	heart primordium	74	25	12.14	0.000184018	0.0066187
ZFA:0000114	heart	292	76	47.89	0.000198928	0.006931388
ZFA:0009052	<i>motor neuron</i>	39	16	6.4	0.000217046	0.007333517
ZFA:0001161	pectoral fin	1019	212	167.12	0.000236558	0.007757698
ZFA:0000041	mesoderm	1309	284	214.68	0.000252216	0.008034876
ZFA:0001185	periderm	344	81	56.42	0.00029706	0.009200606
ZFA:0000135	notochord	612	131	100.37	0.000446455	0.013453969
ZFA:0007001	<i>dorso-rostral cluster</i>	14	8	2.3	0.00060622	0.017322355
ZFA:0007003	<i>ventro-caudal cluster</i>	14	8	2.3	0.00060622	0.017322355
ZFA:0001064	<i>rhombomere</i>	260	63	42.64	0.00062143	0.017322355
ZFA:0009159	mucus secreting cell	77	24	12.63	0.00093149	0.025331978
ZFA:0000304	<i>ventral telencephalon</i>	48	17	7.87	0.001056319	0.028042744
ZFA:0007002	<i>ventro-rostral cluster</i>	15	8	2.46	0.001113075	0.028862289

ZFA:0000470	<b>preoptic area</b>	47	19	7.71	0.001816051	0.046020375
ZFA:0000213	<b>habenula</b>	59	19	9.68	0.002015681	0.049944099
ZFA:0000543	<b>medial longitudinal fasciculus</b>	13	7	2.13	0.002157047	0.052284948
ZFA:0000133	<b>neural rod</b>	249	58	40.84	0.002681813	0.063621744
ZFA:0000609	<b>ventrolateral nucleus</b>	5	4	0.82	0.003132052	0.071270155
ZFA:0001063	posterior caudal vein	5	4	0.82	0.003132052	0.071270155
ZFA:0000113	head mesenchyme	126	33	20.66	0.003259781	0.072693124
ZFA:0000056	pharynx	236	55	38.71	0.003369758	0.07367216
ZFA:0000032	<b>hypothalamus</b>	197	50	32.31	0.003467464	0.07435043
ZFA:0000150	pronephric duct	526	109	86.27	0.004116624	0.084506255
ZFA:0000512	<b>facial lobe</b>	8	5	1.31	0.004263926	0.084506255
ZFA:0001204	axial mesoderm	429	91	70.36	0.004319676	0.084506255
ZFA:0000111	germ ring	213	50	34.93	0.004330285	0.084506255
ZFA:0001000	mesenchyme pectoral fin	29	11	4.76	0.004347542	0.084506255
ZFA:0000338	<b>diencephalic tract/commissure</b>	33	12	5.41	0.004395841	0.084506255
ZFA:0000093	blastomere	105	28	17.22	0.004908621	0.092764613
ZFA:0000086	EVL	154	38	25.26	0.005073941	0.094290735
ZFA:0001306	pharyngeal arch	1013	195	166.14	0.005594287	0.102256227
ZFA:0000045	<b>neural crest</b>	234	64	38.38	0.005694978	0.102417748
ZFA:0000458	<b>ventral thalamus</b>	30	11	4.92	0.005865477	0.103809625
ZFA:0000077	tail bud	469	97	76.92	0.007006635	0.12206872
ZFA:0000944	<b>posterior lateral line</b>	23	9	3.77	0.00764551	0.125569966
ZFA:0001555	<b>epibranchial ganglion</b>	23	9	3.77	0.00764551	0.125569966
ZFA:0001176	blastoderm	198	52	32.47	0.007761719	0.125569966
ZFA:0000188	<b>corpus cerebelli</b>	31	11	5.08	0.007770697	0.125569966
ZFA:0000471	atrium	31	11	5.08	0.007770697	0.125569966
ZFA:0000260	<b>periventricular nucleus</b>	6	4	0.98	0.008174082	0.12791569
ZFA:0000480	<b>caudal octavolateralis nucleus</b>	6	4	0.98	0.008174082	0.12791569
ZFA:0000248	<b>magnocellular preoptic nucleus</b>	9	5	1.48	0.008304633	0.12791569
ZFA:0000138	<b>otic placode</b>	299	65	49.04	0.008438836	0.12791569
ZFA:0001206	intermediate mesoderm	71	20	11.64	0.008542078	0.12791569
ZFA:0000050	optic vesicle	396	83	64.95	0.008604194	0.12791569
ZFA:0000243	<b>neuromast</b>	134	33	21.98	0.008892619	0.130464077
ZFA:0000217	<b>inner ear</b>	76	21	12.46	0.009033161	0.130711666
ZFA:0009073	<b>glial cell</b>	16	7	2.62	0.009143955	0.130711666
ZFA:0009242	<b>primary neuron</b>	100	26	16.4	0.009299871	0.131257668
ZFA:0000082	vein	191	47	31.33	0.009809186	0.136039083
ZFA:0001308	organism subdivision	3372	735	553.03	0.009978545	0.136039083
ZFA:0000083	ventral mesoderm	296	64	48.55	0.010054085	0.136039083
ZFA:0000940	<b>posterior lateral line neuromast</b>	20	8	3.28	0.010126676	0.136039083
ZFA:0001439	anatomical system	4218	877	691.78	0.011091657	0.146235692
ZFA:0001202	optic cup	126	31	20.66	0.011148012	0.146235692
ZFA:0001424	chondrocranium	37	12	6.07	0.012204877	0.156418824
ZFA:0009067	<b>CNS neuron (sensu Vertebrata)</b>	37	12	6.07	0.012204877	0.156418824
ZFA:0000496	compound organ	3445	737	565	0.012864618	0.162049666
ZFA:0000035	lens	321	68	52.65	0.012975034	0.162049666
ZFA:0001391	<b>anterior lateral line ganglion</b>	17	7	2.79	0.013381062	0.162049666
ZFA:0000137	<b>optic stalk</b>	60	17	9.84	0.013819568	0.162049666
ZFA:0000641	<b>cranial nerve</b>	25	9	4.1	0.014051947	0.162049666
ZFA:0000653	<b>dorsal thalamus</b>	25	9	4.1	0.014051947	0.162049666
ZFA:0007009	<b>nerve</b>	25	9	4.1	0.014051947	0.162049666
ZFA:0000164	ventral mesenchyme	51	15	8.36	0.014203479	0.162049666
ZFA:0000344	<b>middle lateral line</b>	10	5	1.64	0.014388266	0.162049666
ZFA:0000939	<b>middle lateral line neuromast</b>	10	5	1.64	0.014388266	0.162049666

ZFA:0005114	<b>middle lateral line system</b>	10	5	1.64	0.014388266	0.162049666
ZFA:0009091	melanocyte	10	5	1.64	0.014388266	0.162049666
ZFA:0007031	<b>anterior neural rod</b>	70	19	11.48	0.015328439	0.165639526
ZFA:0000597	<i>telencephalic tract/commissure</i>	4	3	0.66	0.015449785	0.165639526
ZFA:0001108	<i>anterior commissure</i>	4	3	0.66	0.015449785	0.165639526
ZFA:0009285	podocyte	4	3	0.66	0.015449785	0.165639526
ZFA:0009318	<b>retinal bipolar neuron</b>	4	3	0.66	0.015449785	0.165639526
ZFA:0000038	margin	170	39	27.88	0.01581427	0.167932485
ZFA:0000459	<i>ventromedial thalamic nucleus</i>	7	4	1.15	0.016608115	0.169486707
ZFA:0000578	ceratohyal bone	7	4	1.15	0.016608115	0.169486707
ZFA:0001262	gonad primordium	7	4	1.15	0.016608115	0.169486707
ZFA:0009315	<i>horizontal cell</i>	7	4	1.15	0.016608115	0.169486707
ZFA:0000117	hypoblast	150	35	24.6	0.016720662	0.169486707
ZFA:0001085	hypaxial myotome region	90	23	14.76	0.017055522	0.171323486
ZFA:0000307	<i>vestibulolateralis lobe</i>	14	6	2.3	0.017591	0.175124688
ZFA:0009310	<b>retinal ganglion cell</b>	26	9	4.26	0.01843847	0.181937118
ZFA:0000545	<i>medulla oblongata</i>	70	22	11.48	0.018654509	0.182454187
ZFA:0001291	<i>facial ganglion</i>	18	7	2.95	0.018858834	0.182848694
ZFA:0009051	<i>interneuron</i>	53	15	8.69	0.020246704	0.194612719
ZFA:0001289	<b>ciliary marginal zone</b>	44	13	7.22	0.020846498	0.19866534

## 5.7 Supplementary tables

Supplementary tables can be downloaded at:

[http://bioinfo.unil.ch/supdata/these\\_Julien/Nervous\\_sup\\_dataset.xls](http://bioinfo.unil.ch/supdata/these_Julien/Nervous_sup_dataset.xls)

## 5.8 References

1. Lynch M, Conery JS (2000) The Evolutionary Fate and Consequences of Duplicate Genes. *Science* 290: 1151-1155.
2. Van de Peer Y, Maere S, Meyer A (2009) The evolutionary significance of ancient genome duplications. *Nat Rev Genet* 10: 725-732.
3. Nakatani Y, Takeda H, Kohara Y, Morishita S (2007) Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res* 17: 1254-1265.
4. Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, et al. (2008) The amphioxus genome and the evolution of the chordate karyotype. *Nature* 453: 1064-1071.
5. Jaillon O, Aury J-M, Brunet F, Petit J-L, Stange-Thomann N, et al. (2004) Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* 431: 946-957.
6. Brunet FG, Crollius HR, Paris M, Aury JM, Gibert P, et al. (2006) Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol Biol Evol* 23: 1808-1816.
7. Davis JC, Petrov DA (2004) Preferential duplication of conserved proteins in eukaryotic genomes. *PLoS Biol* 2: e55.
8. Kassahn KS, Dang VT, Wilkins SJ, Perkins AC, Ragan MA (2009) Evolution of gene function and regulatory control after whole-genome duplication: Comparative analyses in vertebrates. *Genome Research* 19: 1404-1418.
9. Amoutzias GD, He Y, Gordon J, Mossialos D, Oliver SG, et al. (2009) Posttranslational regulation impacts the fate of duplicated genes. *Proc Natl Acad Sci USA*.



10. He X, Zhang J (2005) Gene complexity and gene duplicability. *Curr Biol* 15: 1016-1021.
11. Roux J, Robinson-Rechavi M (2008) Developmental Constraints on Vertebrate Genome Evolution. *PLoS Genet* 4: e1000311.
12. Drummond DA, Wilke CO (2008) Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. *Cell* 134: 341-352.
13. Smith CM, Finger JH, Hayamizu TF, McCright IJ, Eppig JT, et al. (2007) The mouse Gene Expression Database (GXD): 2007 update. *Nucl Acids Res* 35: D618-623.
14. Bastian F, Parmentier G, Roux J, Moretti S, Laudet V, et al. (2008) Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species. *in DILS: Data Integration in the Life Sciences*. pp. 124-131.
15. Aitken S (2005) Formalizing concepts of species, sex and developmental stage in anatomical ontologies. *Bioinformatics* 21: 2773-2779.
16. Bard JBL, Kaufman MH, Dubreuil C, Brune RM, Burger A, et al. (1998) An internet-accessible database of mouse developmental anatomy based on a systematic nomenclature. *Mech Dev* 74: 111-120.
17. Smedley D, Haider S, Ballester B, Holland R, London D, et al. (2009) BioMart--biological queries made easy. *BMC Genomics* 10: 22.
18. Sprague J, Bayraktaroglu L, Clements D, Conlin T, Fashena D, et al. (2006) The Zebrafish Information Network: the zebrafish model organism database. *Nucl Acids Res* 34: D581-585.
19. Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, et al. (2009) Ensembl 2009. *Nucl Acids Res* 37: D690-697.
20. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, et al. (2009) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res* 19: 327-335.
21. Penel S, Arigon A-M, Dufayard J-F, Sertier A-S, Daubin V, et al. (2009) Databases of homologous gene families for comparative genomics. *BMC Bioinformatics* 10: S3.
22. Dufayard J-F, Duret L, Penel S, Gouy M, Rechenmann F, et al. (2005) Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics* 21: 2596-2603.
23. R Development Core Team (2007) R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.
24. Alexa A, Rahnenfuhrer J, Lengauer T (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* 22: 1600-1607.
25. Gentleman R, Carey V, Bates D, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5: R80.
26. Smith B, Ashburner M, Rosse C, Bard J, Bug W, et al. (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol* 25: 1251-1255.
27. Akashi H (1994) Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* 136: 927-935.
28. Eppig JT, Bult CJ, Kadin JA, Richardson JE, Blake JA, et al. (2005) The Mouse Genome Database (MGD): from genes to mice--a community resource for mouse biology. *Nucl Acids Res* 33: D471-475.
29. Liao BY, Zhang J (2007) Mouse duplicate genes are as essential as singletons. *Trends Genet* 23: 378-381.

30. Liang H, Li WH (2007) Gene essentiality, gene duplicability and protein connectivity in human and mouse. *Trends Genet* 23: 375-378.
31. Makino T, Hokamp K, McLysaght A (2009) The complex relationship of gene duplication and essentiality. *Trends Genet* 25: 152-155
32. Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A, et al. (2007) Genome-wide atlas of gene expression in the adult mouse brain. *Nature* 445: 168-176.
33. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25: 25-29.
34. Yon Rhee S, Wood V, Dolinski K, Draghici S (2008) Use and misuse of the gene ontology annotations. *Nat Rev Genet* 9: 509-515.
35. Grossmann S, Bauer S, Robinson PN, Vingron M (2007) Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis. *Bioinformatics* 23: 3024-3031.
36. Falcon S, Gentleman R (2007) Using GOSTats to test gene lists for GO term association. *Bioinformatics* 23: 257-258.
37. Davis JC, Petrov DA (2005) Do disparate mechanisms of duplication add similar genes to the genome? *Trends Genet* 21: 548-551.
38. Yeo GW, Van Nostrand E, Holste D, Poggio T, Burge CB (2005) Identification and analysis of alternative splicing events conserved in human and mouse. *Proc Natl Acad Sci USA* 102: 2850-2855.
39. Kopelman NM, Lancet D, Yanai I (2005) Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms. *Nat Genet* 37: 588-589.
40. Su Z, Wang J, Yu J, Huang X, Gu X (2006) Evolution of alternative splicing after gene duplication. *Genome Res* 16: 182-189.
41. Gu X, Su Z (2007) Tissue-driven hypothesis of genomic evolution and sequence-expression correlations. *Proc Natl Acad Sci USA* 104: 2779-2784.
42. Duret L, Mouchiroud D (2000) Determinants of Substitution Rates in Mammalian Genes: Expression Pattern Affects Selection Intensity but Not Mutation Rate. *Mol Biol Evol* 17: 68-70.
43. Christley S, Alber MS, Newman SA (2007) Patterns of Mesenchymal Condensation in a Multiscale, Discrete Stochastic Model. *PLoS Comput Biol* 3: e76.
44. Chamary JV, Parmley JL, Hurst LD (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet* 7: 98-108.
45. Qian W, Zhang J (2008) Gene Dosage and Gene Duplicability. *Genetics* 179: 2319-2324.
46. Gout J-F, Duret L, Kahn D (2009) Differential Retention of Metabolic Genes Following Whole-Genome Duplication. *Mol Biol Evol* 26: 1067-1072.
47. Drummond DA, Wilke CO (2009) The evolutionary consequences of erroneous protein synthesis. *Nat Rev Genet* 10: 715-724.
48. Force A, Lynch M, Pickett FB, Amores A, Yan YL, et al. (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151: 1531-1545.
49. Ohno S (1970) *Evolution by gene duplication*: Springer-Verlag. 160 p.
50. He X, Zhang J (2005) Rapid Subfunctionalization Accompanied by Prolonged and Substantial Neofunctionalization in Duplicate Gene Evolution. *Genetics* 169: 1157-1164.
51. Milinkovitch MC, Helaers R, Tzika AC (2010) Historical Constraints on Vertebrate Genome Evolution. *Genome Biol Evol* 2010: 13-18.

52. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, et al. (2006) In vivo enhancer analysis of human conserved non-coding sequences. *Nature* 444: 499-502.
53. Wapinski I, Pfeffer A, Friedman N, Regev A (2007) Natural history and evolutionary principles of gene duplication in fungi. *Nature* 449: 54-61.
54. Lynch M (2007) *The Origins of Genome Architecture*: Sinauer Associates Inc. 340 p.
55. Coufal NG, Garcia-Perez JL, Peng GE, Yeo GW, Mu Y, et al. (2009) L1 retrotransposition in human neural progenitor cells. *Nature* 460: 1127-1131.
56. Pal C, Papp B, Lercher MJ (2006) An integrated view of protein evolution. *Nat Rev Genet* 7: 337-348.
57. Holland LZ (2009) Chordate roots of the vertebrate nervous system: expanding the molecular toolkit. *Nat Rev Neurosci* 10: 736-746.



# Outlook

---

The work presented in this thesis focused on linking classical models of evo-devo with patterns of genome evolution in model species of vertebrates, mainly human, mouse and zebrafish. I hope that it will contribute to improving the understanding of the role that anatomy and development play in the evolution of genomes and transcriptomes of vertebrates.

My most significant finding was to show, using microarray and EST data spanning zebrafish and mouse development, that developmental processes affect strongly the action of natural selection on animal genomes. Genes expressed early in development tend to be more essential than genes expressed later in development; they also interact more with other genes and are less easily duplicated. These patterns are all consistent with a progressive reduction of evolutionary constraints acting on genes through development. This result was unexpected because morphology of vertebrate species is more conserved at a mid-time point in development than earlier and later in development (i.e. at the “phylotypic” period, the beginning of organogenesis). It was previously suggested that such a conservation of morphology might be the consequence of a maximum of genetic constraints at this period of development. This study taught us that this is not the case and that constraints on the genome do not translate smoothly to other levels of organization such as morphology.

Together with a Master student, Aurélie Comte, we conducted a follow-up study, focusing on searching for markers at the molecular level of the “phylotypic” period. In particular we tested one hypothesis suggesting that the resistance to evolutionary changes of the morphology at this period is due to a high level of interactions. We showed that this hypothesis is not compatible with patterns of protein-protein interactions, signal transduction cascades, and miRNAs expression over the course of zebrafish and mouse development.

Several other analyses are currently in preparation for publication from my PhD work. They also highlight the patterns of evolutionary constraints and opportunities experienced by genomes in the context of anatomy and development. The analyses I could perform during my PhD studies were made possible by the development of an innovative resource, the database Bgee (<http://bgee.unil.ch>), allowing the comparison of

gene expression between different animal species (human, mouse, xenopus, zebrafish, fruitfly). I developed and maintain the pipeline used to gather and integrate multiple sources of expression data (microarrays, ESTs and *in situ* hybridizations) through the use of rigorous statistical analyses. This provides a unique dataset to the evo-devo community and an effort is placed in keeping it up-to-date, based on reference databases such as Ensembl for gene models, the OBO-Foundry for formal descriptions of anatomy and development, ArrayExpress for microarray data, and several model organism databases (ZFIN, MGI, Xenbase, FlyBase) for *in situ* hybridization data.

In a related theoretical work, I focused on the criterion used to decide which anatomical structures can be compared between species. Currently in Bgee, it is possible to compare “homologous” structures, deriving from a common structure in the ancestor of the two species. The comparison of gene expression between such structures is likely to make sense. But homology and its related concepts (e.g. analogy, deep homology) are very complex and debated, particularly when they involve comparisons of evolutionary distant species, for example arthropods and vertebrates. Bgee aims at allowing the comparisons of gene expression in such species, as this can provide answers to long-standing questions regarding homology. For example the evolution of gene expression and its dynamics could be compared between historical homologous structures and structures that are not homologous but functionally equivalent, to know if evolutionary history or function constrain gene expression evolution more. In this context, the different concepts related to homology need to be accommodated rigorously. We gathered their different definitions and developed a formal hierarchical representation to define and organize the concepts discussed in the evolutionary biology literature. The implementation of this ontology to extend the functionalities of Bgee is ongoing.

Thus my PhD work fills some gaps regarding our understanding of the action of selection on animal genomes and transcriptomes. It is also important regarding the development of resources for the community, a necessary step to catalyze many research efforts. The development of resources is particularly timely because of the massive amount of data to come in the next few years with the development of high-throughput sequencing technologies. The integration of such datasets will be complex, but will be a necessary step to access to the promises of such technologies and open new

avenues for the field of inter-species comparisons of gene expression levels. I hope that we will soon be able to move beyond small-scale case studies to examine in a systematic way, and at an unprecedented resolution, the evolutionary specificities of animal genomes. Ultimately, these efforts should bring exceptional insights into the evolution of complex genotype to phenotype maps in animals.





## Appendix 1

Description of the different steps of the pipeline for update of the data of the database Bgee, extracted from the wiki of the lab.

### Update Bgee data

From Mrrwiki (Marc Robinson-Rechavi's lab wiki)

#### Contents

- [1 Introduction](#)
- [2 Before everything](#)
- [3 Ontologies](#)
  - [3.1 Checklist](#)
  - [3.2 Pipeline](#)
  - [3.3 Remarks/Warnings](#)
- [4 Genes, cross-references, gene families](#)
  - [4.1 Checklist](#)
  - [4.2 Pipeline](#)
  - [4.3 Remarks/Warnings](#)
- [5 Affymetrix](#)
  - [5.1 Checklist](#)
  - [5.2 Potential problems](#)
  - [5.3 Pipeline](#)
  - [5.4 Remarks/Warnings](#)
- [6 ESTs](#)
  - [6.1 Checklist](#)
  - [6.2 Pipeline](#)
  - [6.3 Remarks/Warnings](#)
- [7 In-situ](#)
  - [7.1 Checklist zebrafish](#)
  - [7.2 Checklist mouse](#)
  - [7.3 Checklist drosophila](#)
  - [7.4 Pipeline](#)
  - [7.5 Remarks/Warnings](#)
- [8 Differential Expression Affymetrix](#)
  - [8.1 Checklist](#)
  - [8.2 Pipeline](#)
  - [8.3 Remarks/Warnings](#)
- [9 Save a final dump](#)

#### Introduction

This document is here to make future updates of Bgee easier. This checklist will help to keep Bgee up-to-date. Many sources of data are present in Bgee, taken from different sources and databases. This increases the risk of problem/bugs/errors when running an automated pipeline. Here are the different steps to follow.

## Before everything

**This can be done locally.**

- `svn` update to get the last version of the project.
- Keep a stable (frozen) release as a separated branch.  
`svn copy https://svn.vital-it.ch/svn/BGee/trunk/bgee/ https://svn.vital-it.ch/svn/BGee/tags/bgee_v07/`
- Edit `bgee/extra/sql/bgee.sql` with new bgee database name (e.g. `bgee_v07`).
- Edit `bgee/extra/pipeline/pipeline/script_global.sh` with new bgee database name (e.g. `bgee_v07`)

## Ontologies

### Checklist

**This can be done locally.**

- Download the latest version of the anatomy/development ontologies from OBO foundry and put them in `bgee/extra/pipeline/OBO_ontologies/`
  - [Mouse adult gross anatomy](#)
  - [Zebrafish anatomy](#)
  - [Xenopus Anatomy](#)
  - [Drosophila melanogaster](#)
- These ontologies are usually not updated
  - [Mouse anatomy modified by Aitken](#)
  - [Human anatomy modified by Aitken](#)
  - [Human adult anatomy eVoc 2.7](#)
- Check that they are in UNIX format (end-of-line is `\n`). Use `od -c filename` for example. Usually only xenopus ontology uses `\r\n`. Use `"perl -p -e 's/(\r\n|\n|\r)/\n/g'"` if not.
- Check that the `.obo` file is ending with a blank line (specification of OBO format).
- If major changes have been made to an ontology, check its corresponding `.pattern` file.
- Check that no change was made on developmental stages. If it is the case, modify the files `bgee/download/stages.obo` and `bgee/download/stage_association.obo`. Be careful that in `stages.obo`, the stages have to be ordered since it is the only way to create the interval representation and display them in the right order in Bgee.
- Modify `insert_modifs.sql` if some changes have to be made on the ontologies (start/end stages, 'unknown' organ, etc)
- Download the latest version of the [Gene Ontology](#).

**svn commit if this part was done locally.**

### Pipeline

**This is done on devbioinfo server. The project is in `/var/bgee/`**

Go to `bgee/extra/pipeline/pipeline`:

- Edit `bgee/extra/pipeline/pipeline/db_connections.pl` with the new bgee database name (`bgee_v07`).
- Initialize the new database  
`mysql -u root -p < ../../sql/bgee.sql`  
`mysql -u root -p bgee_v07 < ../../sql/bgeeConstraint.sql`  
Don't initialize the `bgeeAdministration` database (conserved from release to release)!
- Insert species infos (mouse, human, zebrafish, xenopus)  
`perl insert_species.pl`
- Insert metastages  
`perl insert_metastages.pl ../../OBO_ontologies/meta_stages.pattern`
- Insert stages infos from OBO files

- ```
perl insert_stages.pl ../OBO_ontologies/xenopus_stages.pattern
perl insert_stages.pl ../OBO_ontologies/mouse_stages.pattern
perl insert_stages.pl ../OBO_ontologies/human_stages.pattern
perl insert_stages.pl ../OBO_ontologies/zebrafish_stages.pattern
perl insert_stages.pl ../OBO_ontologies/fly_development.pattern
```
- Insert organs infos from OBO files

```
perl insert_organs.pl ../OBO_ontologies/adult_mouse_anatomy.pattern
perl insert_organs.pl ../OBO_ontologies/context-plus-mouse.pattern
perl insert_organs.pl ../OBO_ontologies/evoc_anatomicalsystem.pattern
perl insert_organs.pl ../OBO_ontologies/context-plus-human.pattern
perl insert_organs.pl ../OBO_ontologies/xenopus_anatomy.pattern
perl insert_organs.pl ../OBO_ontologies/zebrafish_anatomy.pattern
perl insert_organs.pl ../OBO_ontologies/fly_anatomy.pattern
perl insert_organDescendants.pl
```
  - Insert gene ontology annotations

```
perl insert_go_terms.pl
```
- Remarks/Warnings**
- [Inferred relationships](#)
  - Warning with `insert_organs.pl` when inferring the new relationships (the same relation is sometimes inferred several times from different paths).

```
DBD::mysql::st execute failed: Duplicate entry 'ZFA:0009000-ZFA:0009082' for
key 1 at insert_organs.pl line 387.
DBD::mysql::st execute failed: Duplicate entry 'ZFA:0009000-ZFA:0009082' for
key 1 at insert_organs.pl line 387.
```

TO DO: Change the array with a hash to suppress this problem.
  - Warning with drosophila organs due to problems in the OBO file. I reported them to the OBOfoundry tracker but they are not yet corrected.
    - FBbt:00002779 has 2 identical synonyms
    - FBbt:00004158 is\_a FBbt:00004159, but also part\_of...

## Genes, cross-references, gene families

### Checklist

**This is done on devbioinfo server.**

- Update Ensembl perl API on the server to the new version. See [instructions](#) for details... Or ask Sebastien.
- Check that biomart is updated to the new Ensembl version (Either on [Biomart](#) or [Ensembl](#) server) and edit `bgee/extra/pipeline/pipeline/db_connections.pl` accordingly (in subroutine `send_biomart_query`).

### Pipeline

- Insert all infos on genes

```
perl insert_genes.pl xenopus
perl insert_genes.pl zebrafish
perl insert_genes.pl mouse
perl insert_genes.pl human
perl insert_genes.pl fruitfly
perl insert_geneName_zebrafish_miRNA.pl
perl insert_xref_zebrafish_miRNA.pl
```
- Insert homology groups (Ensembl families and Gene Trees orthologs). The gene trees are inserted with the taxonomic depth indicated as argument (117571=Euteleostomi, 33213=Bilateria and includes drosophila)

```
perl insert_families.pl
perl insert_orthologs.pl 117571
perl insert_orthologs.pl 33213
```

- ```
perl insert_miRNA_families.pl
```
- Fill geneToTerm table  
`perl insert_geneToTerm.pl` (takes 12H)
  - Insert ponctual modifications  
`mysql -u root -p bgee_v07 < insert_modifs.sql`
  - If for some genes, the field "geneBioTypeId" is NULL:  
`INSERT INTO geneBioType (geneBioTypeName) VALUE ('unknown');`  
`UPDATE gene SET geneBioTypeId = XX WHERE geneBioTypeId IS NULL;`
  - Save an intermediate dump  
`mysqldump -u root -p --no-create-info --no-create-db --triggers=false bgee_v07 > dump_genes.sql`

#### Remarks/Warnings

- In `insert_genes.pl`, the URL where we download the cross-links to 4DXpress depends on the date it was created (e.g. [http://4dx.embl.de/bgee/ensIds\\_2009\\_03.txt](http://4dx.embl.de/bgee/ensIds_2009_03.txt)). Frederic asked Thorsten to change that: we'll have to modify the script when it's done.
- There can be sometimes problems with the API connection that is lost (problem on Ensembl side). Just rerun the script, it usually works fine.  
`DBD::mysql::st execute failed: Lost connection to MySQL server during query at /usr/local/ensembl/ensembl-compara/modules/Bio/Ensembl/Compara/DBSQL/MemberAdaptor.pm line 662.`
- In `insert_geneToTerm.pl` teher can be some duplicate entriy errors because some terms differ just by upper/lower case.
- In `insert_orthologs.pl`, warning of perl when the recursion is going deep:  
`Deep recursion on subroutine "Bio::Ensembl::Compara::NestedSet::_recursive_get_all_leaves" at /usr/local/Ensembl/ensembl-compara/modules/Bio/Ensembl/Compara/NestedSet.pm line 1508.`

## Affymetrix

### Checklist

- Convert `annotation.xls` to text files `affymetrixChip`, `microarrayExperiment` and `chipType` (if doing this from Excel, be careful with the end-of-line characters).
    - Put them in `bgee/extra/pipeline/curation/Affymetrix/`
  - Raw data verification (`.cel` files)
    - Download the latest version of the folder  
`bgee/extra/pipeline/curation/Affymetrix/cel_data/` on the annotators computers.
  - Raw data with only 1 `.cel` file
    - It's not possible in that case to use `gcRMA`
    - Normalization with `MAS5`: use  
`bgee/extra/pipeline/Affymetrix/bioconductor/affy_analysis_mas5.R`
  - Processed data split and verification (`MAS5`)
    - Download the latest version of the folder  
`bgee/extra/pipeline/curation/Affymetrix/processed_mas5/` on the annotators computers.
    - Put them in the `bgee/extra/pipeline/Affymetrix/processed_mas5/not_separated/` folder.
    - In `bgee/extra/pipeline/pipeline/:`
- ```
perl separate_affy_processed_mas5.pl
```
- Be careful, some processed data have a blank column instead of probeset IDs. This is frequent and is due to `ArrayExpress`. The experiments should be removed and put in "not\_included for now" (or the probeset

column should be taken from somewhere else, because we know the chip type, but no guarantee that the probesets are in the same order, so it's dirty).

- In `bgee/extra/pipeline/pipeline/` check that there is no problem left with the annotation:  
`perl check_affy_curation.pl` before (before normalization, to detect problems in the annotation and files. Be careful, the script should be able to connect to the database, and it should be run on the computer having all the data. It checks a lot of small common mistakes by the annotators)
- Raw data normalization
  - Send these folders to vital-IT/devbioinfo using `rsync`. For example:  

```
rsync -Wav -essh --exclude '*.gz'
~/work/bgee/extra/pipeline/Affymetrix/cel_data/ jroux@dee-serv02.vital-
it.ch:/scratch/frt/yearly/jroux/pipeline/Affymetrix/cel_data/
rsync -Wav -essh ~/work/bgee/extra/pipeline/Affymetrix/bioconductor/
jroux@dee-serv02.vital-
it.ch:/scratch/frt/yearly/jroux/pipeline/Affymetrix/bioconductor/
rsync -Wv -essh ~/work/bgee/extra/pipeline/Affymetrix/* jroux@dee-
serv02.vital-it.ch:/scratch/frt/yearly/jroux/pipeline/Affymetrix/
```
  - Send also the last version of the scripts to Vital-IT. For example:  

```
rsync -Wv -essh ~/work/bgee/extra/pipeline/pipeline/*.pl jroux@dee-
serv02.vital-it.ch:/scratch/frt/yearly/jroux/pipeline/pipeline/
```

Or `svn update` on devbioinfo.
  - Connect to vital-IT  

```
cd /scratch/frt/yearly/jroux/pipeline/pipeline/ or alias bgee
mkdir /scratch/frt/yearly/jroux/pipeline/Affymetrix/processed_schuster/
perl launch_affy_analysis.pl (or perl launch_affy_analysis.pl EXP_ID if you want to
analyse only one experiment).
```
  - When calculations are finished, download the data produced:  

```
rsync -Wav -essh /scratch/frt/yearly/jroux/pipeline/Affymetrix/bioconductor/
admin@130.223.48.XX:~/work/bgee/extra/pipeline/Affymetrix/bioconductor/
rsync -Wav -essh
/scratch/frt/yearly/jroux/pipeline/Affymetrix/processed_schuster/
admin@130.223.48.XX:~/work/bgee/extra/pipeline/Affymetrix/processed_schuster
/
```
- In `bgee/extra/pipeline/pipeline/` check that there is no problem with the normalized files:  
`perl check_affy_curation.pl` after (after normalization)
- You can compress the `cel` files already used. In the subdirectory `cel_data/`:  
`find . ! -name '*.gz' -type f -exec gzip -9 {} \;`

### Potential problems

- When normalizing:
  - The annotation package for the chipType is missing and it can't be automatically installed on vital-IT (no permission) -> ask Li Long to install it. The List of all annotations packages is [here](#). Another option is to do the normalization on another computer (yours or devbioinfo) where you have the rights to install the required packages.
  - The annotation file asks you to normalize an experiment with only 1 chip. This is not possible with `gcrma` -> has to be done with `mas5` (use `bgee/extra/pipeline/Affymetrix/bioconductor/affy_analysis_mas5.R`). Put the results in `processed_mas5/` folder and change the normalization and detection method in `affymetrixChip` (usually from 2 and 2 to 1 and 1). Examples of error messages:  

```
> data.gcrma <- gcrma(data, affinity.info=ai, type="affinities")
Adjusting for optical effect.Done. Error in model.frame(formula,
rownames, variables, varnames, extras, extranames, :
variable
lengths differ (found for 'x') > data.gcrma <- gcrma(data,
affinity.info=ai, type="affinities") Adjusting for optical
effect.Done. Error in model.frame.default(formula = y ~ x,
```

```
drop.unused.levels = TRUE) : variable lengths differ (found for 'x')
Calls: gcrma ... lm -> eval -> eval -> model.frame ->
model.frame.default
```

- Annotations errors. A problem happens if the annotators put the wrong chip type. This is often the case when multiple chip types are used in the same experiment. Sometimes the mistake is present in ArrayExpress! -> You can have a look at the chip type directly in the header of the CEL file.
- Corrupted files. Usually there is nothing to do... You can remove the problematic chip from the experiment in the annotation file.
- Custom chips: the annotation package does not exist in bioconductor (encode chips for example)
- Memory problems: shouldn't occur on vital-IT (216Go memory)
- Other problems may be solved using the [bioconductor mailing list](#) usually.
- With processed data (processed\_mas5):
  - Annotation problems: the name in the annotation file does not correspond to the name of the files in the experiment folder.
  - Corrupted files: sometimes the exported files on ArrayExpress do not include the probeIds. The first column if the downloaded files is empty -> find the list of probes from another experiment using the same chip, or contact AE so that they correct their file.
  - Corrupted files: sometime the files don't have the correct number of lines (probes). This is mysterious and is probably due to a bad submission to AE.

### Pipeline

- Transfer the processed data on the machine where the pipeline is run
 

```
rsync -Wav -essh extra/pipeline/Affymetrix/processed_schuster/
bgee@devbioinfo:/var/bgee/extra/pipeline/Affymetrix/processed_schuster/
rsync -Wav -essh extra/pipeline/Affymetrix/processed_mas5/
bgee@devbioinfo:/var/bgee/extra/pipeline/Affymetrix/processed_mas5/
```
- Fill the affymetrixProbeset table and noExpressionAffymetrixProbeset table. If you use the "mysql" option it will insert directly into the database, but this is too slow (many days compared to 3hours with "file" for bgee\_v07):
 

```
perl create_files_affy.pl <expr/no_expr/both> <mysql/file>
```
- Sort the files (supposed to be faster to insert primary keys, to check)
 

```
sort -k1 -T /var/tmp affymetrixProbeset.tsv > temp_file
mv temp_file affymetrixProbeset.tsv
sort -k1 -T /var/tmp noExpressionAffymetrixProbeset.tsv > temp_file
mv temp_file noExpressionAffymetrixProbeset.tsv
```

 this is taking ~1 hour.
- Load the files into mysql. This is faster with no constraints so you have to reinitiate the database:
 

```
mysqldump -u root -p --no-create-info --triggers=false bgee_v07 >
dump_v07_affymetrixChip.sql
mysql -u root -p < ../../sql/bgee.sql
mysql -u root -p bgee_v07 < dump_v07_affymetrixChip.sql
mysql -u root -p bgee_v07 -e "load data infile '$PWD/expression.tsv' into
table expression; SHOW WARNINGS"
mysql -u root -p bgee_v07 -e "load data infile '$PWD/noExpression.tsv' into
table noExpression; SHOW WARNINGS"
mysql -u root -p bgee_v07 -e "load data infile
'$PWD/noExpressionAffymetrixProbeset.tsv' into table
noExpressionAffymetrixProbeset; SHOW WARNINGS"
mysql -u root -p bgee_v07 -e "load data infile '$PWD/affymetrixProbeset.tsv'
into table affymetrixProbeset; SHOW WARNINGS"
```

 This is taking more than 1 hour (bgee\_v07)  
 If you want to use nohup, you have to paste the MySQL root password to -p => -pXXXXX in order to use nohup in a non-interactive way!
- Load the constraints



```
mysql -u root -p bgee_v07 < ../../sql/bgeeConstraint.sql
```

This is long! (31 hours with bgee\_v07)

- Save an intermediate dump  

```
mysqldump -u root -p --no-create-info --triggers=false bgee_v07 > dump_affy.sql
```

#### Remarks/Warnings

- Because many folders (for cel files) downloaded in Arrayexpress finish by ".raw" or ".raw.1", you can rename them using:  

```
for i in *.raw; do mv $i ${i/.raw/}; done
```
- Insert one condition only:  

```
perl create_files_affy.pl <expr/no_expr/both> <mysql/file> <organId> <stageId>
```
- To do: script to clean the expression table for 1 organ/stage? (if the insertion of an experiment has gone wrong)  

```
perl update_expression_affy.pl <organId> <stageId> (Warning! script not up to date)
```
- It is always possible to use the old scripts, but they are really long!  

```
perl insert_affy.pl <present/absent/both>
perl check_affy_inserted.pl
perl update_affymetrixProbeset.pl
perl insert_expression_affy.pl <expr/no_expr/both>
```
- Data files (.cel) and MAS5 are stored locally on my computer and are not on the svn -> find a solution?
- Consensus concerning probeset quality (when multiple probesets are present for the same genes):

|          | Pst/High | Pst/Low | Abs/High | Abs/Low |
|----------|----------|---------|----------|---------|
| Pst/High | Pst/High |         |          |         |
| Pst/Low  | Pst/High | Pst/Low |          |         |
| Abs/High | Pst/Low  | Pst/Low | Abs/High |         |
| Abs/Low  | Pst/Low  | Pst/Low | Abs/High | Abs/Low |

## ESTs

### Checklist

**This can be done locally.**

- Annotation files (annotation\_libs\_....txt) are located in  
bgee/extra/pipeline/curation/EST/.
- Download the latest version of [Mm.data](#), [Hs.data](#), [Dr.data](#), [Str.data](#), [Dm.data](#) and uncompress them.
- Download the latest version of [library.report](#).
- Put them in bgee/extra/pipeline/EST\_NCBI/.
- Synchronize them with the server:  

```
rsync -Wav -essh ~/work/bgee/extra/pipeline/EST_NCBI/*.data
bgee@devbioinfo:/var/bgee/extra/pipeline/EST_NCBI/
rsync -Wav -essh ~/work/bgee/extra/pipeline/EST_NCBI/library.report
bgee@devbioinfo:/var/bgee/extra/pipeline/EST_NCBI/
```
- For the script insert\_miRNA\_est.pl:
  - Check that the file organs\_correspondances.csv (manually curated) exists and that it's placed in  
../curation/miRNA/.
  - Download the latest version of [S.xls](#), save it as a tsv (keep the name as S.csv) and place it in  
../miRNA/EST\_smiRNadb/.

- Download the latest version of the files [Report\\_x.csv](#) (check Downloads section) and place them in `../miRNA/EST_smRNAdb/`.
- For the mapping between Flybase genes and UniGene clusters
  - Download `dmel_all_cdna.fasta` from biomart  
Dataset name = "dmelanogaster\_gene\_ensembl"  
Filter (biotype) = "protein\_coding"  
Under the "Sequences" category:  
Attribute = "ensembl\_gene\_id"  
Attribute = "ensembl\_transcript\_id"  
Attribute name = "cDNA sequences"
  - - Download from Unigene the file [Dm.seq.uniq](#)
    - `perl change_fasta_headers.pl Dm.seq.uniq > Dm.seq.uniq_new_headers`
    - `formatdb -p F -i dmel_all_cdna.fasta -n my_db`
    - `blastall -p blastn -F F -m8 -d my_db -i ./Dm.seq.uniq_new_headers -a4 -e 1e-10 -o dmel_cdna.results`
    - `perl extract_results.pl`

### Pipeline

**This is done on devbioinfo server.**

- Insert normal EST libraries and their stage and organ. Indicate the mapping file if it is not available from Biomart.  
`perl insert_est.pl mouse`  
`perl insert_est.pl human`  
`perl insert_est.pl zebrafish`  
`perl insert_est.pl xenopus`  
`perl insert_est.pl fruitfly`  
`../EST_NCBI/mapping/mapping_dmel_unigene_ensembl.txt`  
`perl insert_miRNA_est.pl fruitfly`  
`perl insert_miRNA_est.pl human`  
`perl insert_miRNA_est.pl mouse`  
`perl insert_miRNA_est.pl zebrafish`
- Fill expression table and update the field `estData` (quality) in EST table  
`perl insert_expression_est.pl` (Be careful, this has to be done after the insertion of affymetrix data into expression table).
- Save an intermediate dump  
`mysqldump -u root -p --no-create-info --triggers=false bgee_v07 > dump_EST.sql`

### Remarks/Warnings

- TO DO: script to check that all annotations are corresponding to stages/organs present in Bgee.
- TO DO: script to check that no space has been inserted by mistake (before or after a tab especially) in the annotation file.

## *In-situ*

### Checklist zebrafish

- Update the mapping OBO Ids to ZFIN Ids for organs and stages. In `bgee/extra/pipeline/In_situ/ZFIN/`:
  - `perl mk_organs_correspondance.pl`
  - `perl mk_stages_correspondance.pl`



### Checklist mouse

- Connect to the database MGI with [SquirrelSQL Client](#) (increased memory) or any other client if you manage (good luck!)  
cd /Applications/Squirrel/  
java -Xmx512m -Xms256m -jar squirrel-sql.jar to install SquirrelSQL Client.  
Add Sybase plugin during the installation process  
Unzip the Sybase driver [Media:jconnect60.zip](#)  
Once the installation is complete, run squirrelsqli and select the *Drivers* tab  
Modify the driver *Sybase Adaptive Server Anywhere*:  
Add an *Extra Class Path* to jConnect-6\_0/classes/\* jar (directory previously unzipped)  
Then, click on *List Drivers* and select **com.sybase.jdbc3.jdbc.SybDriver** as *Class Name*.  
Click *OK* to finish. *Sybase Adaptive Server Anywhere* driver should be activated now.  
Go back to *Aliases* tab, and create a new alias with these parameters:  
URL: jdbc:sybase:Tds:gondor.informatics.jax.org:4025/mgd  
user: jroux  
pwd: JR0ux01  
Increase memory for squirrelsqli, tables you will retrieve are very large !
- Load the following tables and save them (right click) into tab delimited .csv files. Be careful to download them on the same day (regular updates are made on MGI).  
Un-select limit rows checkbox !!!  
Save result table as *table\_name.csv*, with *Include column headers*, *Export CSV file*, *Use tab character*, Line separator: LF (n), Charset: UTF-8, and *Export complete table*  
SELECT \* FROM ACC\_Accession WHERE (\_LogicalDB\_key=60 OR \_LogicalDB\_key=83)  
AND \_MGIType\_key=2  
SELECT \* FROM ALL\_Allele  
SELECT \* FROM GXD\_AlleleGenotype  
SELECT \* FROM GXD\_Genotype  
SELECT \* FROM GXD\_Assay  
SELECT \_Specimen\_key, \_Assay\_key, \_Genotype\_key FROM GXD\_Specimen  
SELECT \_Result\_key, \_Specimen\_key, \_Strength\_key, \_Pattern\_key FROM  
GXDI\_nSituResult  
SELECT \* FROM GXD\_ISResultStructure  
SELECT \* FROM GXD\_Structure

### Checklist drosophila

- You have to install the BDGP database
- Download the dump [here](#), put it in bgee/extra/pipeline/In\_situ/BDGP and load it:  
create database exgo\_200703 character set utf8;  
mysql -u root -p exgo\_200703 < exgopub-20070309.dump

### Pipeline

```
perl insert_in_situ_zfin.pl <present/absent/both>
perl insert_in_situ_mgi.pl <present/absent/both>
perl insert_in_situ_bdgp.pl
perl insert_in_situ_xenbase.pl
perl insert_expression_in_situ.pl <expr/no_expr/both> (Be careful, this has to be done
after the insertion of affymetrix data into expression table).
```

### Remarks/Warnings

- Some fields are not downloaded in MGI tables because they contain "\n" and the exported files are corrupted because of that.
- In case of missing data in the publications (e.g. "skeletal muscle" part\_of "hindlimb" or part\_of "forelimb"), MGI curators create a new term in the ontology "skeletal muscle", at the same level than "hindlimb" and "forelimb"

(i.e. children of "limb"). These new terms are not in the Edimburgh ontology (EMAP)... In our logic these data should be mapped to the upper level "limb". For now we didn't integrate these problematic data.

- MGI stores only [normal conditions](#) expression data (wild-type and mutants, but no treatment, etc)
- Quality codes in MGI:

| code | quality        | percentage of the expression results | quality in Bgee |
|------|----------------|--------------------------------------|-----------------|
| -2   | Not Applicable | 0                                    | Not included    |
| -1   | Not Specified  | 0                                    | Not included    |
| 1    | Absent         | 31                                   | Not included    |
| 2    | Present        | 42                                   | High            |
| 3    | Ambiguous      | 1.8                                  | Low             |
| 4    | Trace          | 0.2                                  | Low             |
| 5    | Weak           | 14                                   | Low             |
| 6    | Moderate       | 1.2                                  | High            |
| 7    | Strong         | 10                                   | High            |
| 8    | Very strong    | 0.15                                 | High            |

- Quality codes in ZFIN:

| Thisse stars | quality                                                                                                         | quality in Bgee |
|--------------|-----------------------------------------------------------------------------------------------------------------|-----------------|
| ★            | Probe is difficult to use. General basal level of expression with more intense labeling in particular structure | Low             |
| ★ ★          | Weak expression pattern                                                                                         | Low             |
| ★ ★ ★        | Moderate expression pattern.                                                                                    | High            |
| ★ ★ ★ ★      | Nice strong expression pattern                                                                                  | High            |
| ★ ★ ★ ★ ★    | Simple to use, intense expression pattern restricted to a few structures                                        | High            |
| No star      | (Experiments not made by Thisse)                                                                                | High            |

## Differential Expression Affymetrix

### Checklist

- You can do this analysis on your local computer (not really demanding). Be careful that you have all the data then.
- Launch differential analysis  

```
perl launch_diff_analysis.pl
```
- Send data to the server devbioinfo using rsync:  

```
rsync -Wav -essh
~/work/bgee/extra/pipeline/Affymetrix/processed_differential/
bgee@devbioinfo:/var/bgee/extra/pipeline/Affymetrix/processed_differential/
rsync -Wav -essh ~/work/bgee/extra/pipeline/Affymetrix/bioconductor/
bgee@devbioinfo:/var/bgee/extra/pipeline/Affymetrix/bioconductor/
```

### Pipeline

- Fill the deaAffymetrixProbesetSummary table  

```
perl insert_diff_affy.pl
```
- Fill the differentialExpression table  

```
perl insert_diff_expression_affy.pl
```

### Remarks/Warnings

- It is possible to launch the scripts on one experiment/condition only:  
perl launch\_diff\_analysis.pl <expId>  
perl insert\_diff\_affy.pl <expId> <chipTypeId>  
perl insert\_diff\_expression\_affy.pl <organId> <stageId>
- Data files are stored locally on my computer and are not on the svn -> find a solution?

### Save a final dump

- mysqldump -u root --no-create-info --triggers=false bgee\_v07 > dump\_vXX.sql
- ... and open a bottle to celebrate :p

## Appendix 2

Brochure describing the functionalities of the database Bgee.

### Some questions Bgee can answer

- What is the expression in zebrafish of the 2 orthologs of the mammalian gene *pax6* ?
- Is the expression of the gene *evx1* conserved between vertebrates ?
- How many orthologs are expressed at pharyngula in vertebrates ?
- Which zebrafish organ is the homolog of mammalian lungs ?

### Contact

Marc Robinson-Rechavi's lab

Department of Ecology and Evolution  
University of Lausanne  
Swiss Institute of Bioinformatics

bgee@isb-sib.ch

### Reference

**Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species.**  
Bastian F. et al. (2008) In DILS: Data Integration in Life Sciences. Lectures Notes in Computer Science. 5109:124-131

Image credits:  
*In-situ* hybridization images are under the Creative Commons Attribution License and are adapted from Bertola et al. BMC Evol Bio 2008, 8:166. Other images are adapted from George Shuklin (mouse), the U.S. Geological Survey (xenopus), André Karwath (drosophila) and Guillaume Paumier (microarray).

A database for the study of Gene Expression Evolution

<http://bgee.unil.ch>
<http://bgee.unil.ch>
<http://bgee.unil.ch>

### Presentation

- Bgee is a database to compare gene expression patterns between animals.
- We integrate heterogeneous gene expression data (EST, Affymetrix and *in-situ* hybridization) and link them to anatomy and development.
- We define homology and analogy relationships between anatomies of species.

### Homologous expression of the gene *acta1* in zebrafish and mouse

- Read every paper:  
Figures 6H and 4B, Bertola et al. BMC Evol Biol 2008, 8:166
- Or use Bgee:

Danio rerio

Mus musculus

| Homologous organs common to all the selected species, with expression data | Danio rerio                                                                                                       | Mus musculus                                                                                                      |
|----------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------|
| <a href="#">show/hide genes details</a>                                    | <a href="#">See genes details</a>                                                                                 | <a href="#">See genes details</a>                                                                                 |
|                                                                            | zgc:112098                                                                                                        | Acta1                                                                                                             |
| <input checked="" type="checkbox"/> HOG:0000071: anatomical structure      | 4 est (4 libraries) - 36 probesets (36 chips) - 2 in situ evidences (2 experiments) - Expression in substructures | 388 est (28 libraries) - 134 probesets (134 chips) - Expression in substructures                                  |
| <input checked="" type="checkbox"/> HOG:0000152: embryonic mesoderm        | Expression in substructures                                                                                       | Expression in substructures                                                                                       |
| <input checked="" type="checkbox"/> HOG:0000145: paraxial mesoderm         | 1 in situ evidences (1 experiments)                                                                               | -                                                                                                                 |
| <input checked="" type="checkbox"/> HOG:0000191: somites                   | 2 in situ evidences (2 experiments)                                                                               | 1 in situ evidences (1 experiments)                                                                               |
| <input checked="" type="checkbox"/> HOG:0000302: cardiovascular system     | Expression in substructures                                                                                       | Expression in substructures                                                                                       |
| <input checked="" type="checkbox"/> HOG:0000726: heart                     | 12 est (2 libraries) - 5 probesets (5 chips) - 1 in situ evidences (1 experiments)                                | 6 est (2 libraries) - 32 probesets (32 chips) - 3 in situ evidences (1 experiments) - Expression in substructures |
| <a href="#">...</a>                                                        |                                                                                                                   |                                                                                                                   |

<http://bgee.unil.ch>
<http://bgee.unil.ch>
<http://bgee.unil.ch>

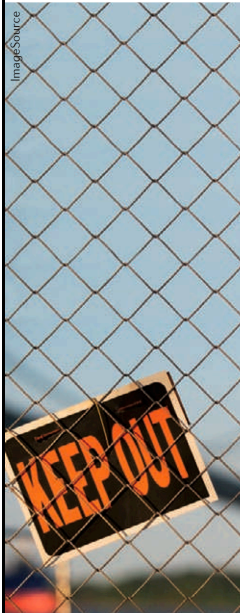
## Appendix 3

Highlight of the article “Developmental Constraints on Vertebrate Genome Evolution” (chapter 3) published in Nature Reviews Genetics.

### RESEARCH HIGHLIGHTS

Nature Reviews Genetics | AOP, published online 13 January 2009; doi:10.1038/nrg2530

 **EVO-DEVO**



## Development sets the limits

It has long been suggested that organismal evolution is limited by constraints on embryonic development that prevent certain changes from being tolerated. A new study using zebrafish and mice has explored how developmental constraints determine which genetic changes can be tolerated and therefore contribute to evolution.

Previous studies have examined developmental constraints on evolutionary genetic change in vertebrates. However, these studies might not give a reliable picture of the pattern of constraint because of issues such as coarse divisions of developmental stages into simply 'early' and 'late', and small sample sizes in terms of the number of genes analysed.

Roux and Robinson-Rechavi first built up gene expression profiles for 26 mouse and 14 zebrafish

developmental stages using existing data on thousands of genes from EST counts and microarray analyses, respectively. They then examined the likelihood that two types of genetic change — loss of function and duplication — are tolerated for genes expressed at each stage.

In terms of loss of function, severe phenotypes reported from mutagenesis and morpholino studies were used to indicate probable evolutionary constraint. In the case of gene duplication, the expression of gene duplicates that have been retained following ancestral genome duplication was examined; the retention of a duplicate suggests that doubling the gene dose has been tolerated during evolution. For both species, a clear trend was seen in which the early stages of development are the least tolerant of mutations.

Following this early peak, constraint seems to decline steadily throughout the rest of development.


Notably, these findings contrast with the 'hourglass' model that is often used to describe vertebrate developmental constraint that is seen at the morphological level. Here, maximal constraint is seen at a mid-developmental time point, before and after which evolution has greater flexibility to act. As the authors point out, our ability to make sense of differences in timing of constraint at the two levels will require an increased understanding of how mutation gives rise to ontogenic change during evolution.

Louisa Flintoft

**ORIGINAL RESEARCH PAPER** Roux, J. & Robinson-Rechavi, M. Developmental constraints on vertebrate genome evolution. *PLoS Genet.* **4**, e100031 (2008)

## Appendix 4

Evaluation of the article “Developmental Constraints on Vertebrate Genome Evolution” (chapter 3) published in Faculty of 1000.

 **Must Read**

F1000 Factor **6.0**

EndNote

Download citation

Send page by email

**Developmental constraints on vertebrate genome evolution.**  
Roux J, Robinson-Rechavi M  
*PLoS Genet* 2008 Dec **4**(12):e1000311 [[abstract on PubMed](#)] [[citations on Google Scholar](#)] [[related articles](#)] [[FREE full text](#)]

**Selected by** | Nicolas Galtier  
Evaluated 20 Feb 2009  
[Relevant Sections](#)

**Faculty Comments & Author Responses**

**Faculty Member**  
**Nicolas Galtier**  
Université Montpellier II,  
France  
Genomics & Genetics  
☒ Confirmation  
☐ New Finding  
☐ Controversial

**Comments**  
**This well-conducted study demonstrates that, in vertebrates, genes involved in early developmental stages are more constrained during evolution than genes expressed late. This is perhaps not a big surprise but for sure worth knowing.**  
  
The main strength of this paper is in the accumulation of independent corroborative arguments: early-expressed genes are enriched in essential genes (whose loss of function is lethal), get back more rapidly to single copy after gene duplication, and experience a lower nonsynonymous/synonymous ratio, on average, than late-expressed genes -- and these properties are detected in two distinct species (mouse and zebrafish). The clear-cut demonstration contrasts with the controversial morphological literature on the subject. The famous "hourglass" hypothesis, which states that the level of constraint reaches a maximum at an intermediate developmental stage, is contradicted here.  
  
**Competing interests:** None declared  
Evaluated 20 Feb 2009  
[How to cite this evaluation](#)

## Appendix 5

Article published in *BMC Genomics* on a project led by Fernando Cruz (post-doc in the lab). My contribution is on the analysis of Gene Ontology enrichment and on the extraction, mapping and normalization of microarray gene expression data.

Research article

Open Access

## The expansion of amino-acid repeats is not associated to adaptive evolution in mammalian genes

Fernando Cruz\*<sup>1,2</sup>, Julien Roux<sup>1,2</sup> and Marc Robinson-Rechavi<sup>1,2</sup>

Address: <sup>1</sup>Department of Ecology and Evolution, Biophore, University of Lausanne, 1015 Lausanne, Switzerland and <sup>2</sup>Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland

Email: Fernando Cruz\* - [Fernando.CruzRodriguez@unil.ch](mailto:Fernando.CruzRodriguez@unil.ch); Julien Roux - [julien.roux@unil.ch](mailto:julien.roux@unil.ch); Marc Robinson-Rechavi - [Marc.Robinson-Rechavi@unil.ch](mailto:Marc.Robinson-Rechavi@unil.ch)

\* Corresponding author

Published: 18 December 2009

Received: 1 September 2009

BMC Genomics 2009, 10:619 doi:10.1186/1471-2164-10-619

Accepted: 18 December 2009

This article is available from: <http://www.biomedcentral.com/1471-2164/10/619>

© 2009 Cruz et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

### Abstract

**Background:** The expansion of amino acid repeats is determined by a high mutation rate and can be increased or limited by selection. It has been suggested that recent expansions could be associated with the potential of adaptation to new environments. In this work, we quantify the strength of this association, as well as the contribution of potential confounding factors.

**Results:** Mammalian positively selected genes have accumulated more recent amino acid repeats than other mammalian genes. However, we found little support for an accelerated evolutionary rate as the main driver for the expansion of amino acid repeats. The most significant predictors of amino acid repeats are gene function and GC content. There is no correlation with expression level.

**Conclusions:** Our analyses show that amino acid repeat expansions are causally independent from protein adaptive evolution in mammalian genomes. Relaxed purifying selection or positive selection do not associate with more or more recent amino acid repeats. Their occurrence is slightly favoured by the sequence context but mainly determined by the molecular function of the gene.

### Background

Microsatellites or simple sequence repeats (SSRs) are DNA tracts composed of 1-6 bp long motifs repeated in tandem. A balance between slippage events, that increase the purity of the repeat, and point mutations, that tend to eliminate perfect repeats, determines their length distribution. However, as the slippage rate is higher than the point mutation rate, the purity of the repeated tract will be an inverse measure of the age of the SSR [1-3].

Triplet repeats are more common within coding regions [4], as they are less likely to alter the reading frame and can be translated into amino-acid repeats (AARs). AARs

are frequently associated with disease [e.g. [5,6]]. Strong effects on morphology and phenotype have also been described in dog breeds [7]. Examples of AARs contributing to adaptive evolution [2,8] have been found in case studies in insects [9], plants [10,11] and mammals [12].

Genomic comparisons have shown that highly variable AARs have a higher purity in their coding sequence [13,14]. AAR expansion has been found to correlate with the non-synonymous rate of substitution [13,15,16] supporting a role of selection in their expansion. The correlation is consistent with either relaxed purifying selection, or with positive selection; the latter is suggested by case

Page 1 of 10

(page number not for citation purposes)



studies of adaptive evolution [9-12]. Previous studies [13,15,16] have been restricted in their taxonomic scale, did not take into account exon boundaries, and did not integrate potential confounding parameters into their analyses. Here we perform a systematic study of mammalian genomes. We contrasted AARs in positively selected genes (PSGs) and non-PSGs [17] to examine their relationship with protein adaptive evolution. We also analyzed other factors correlating with AARs in 6 high coverage mammalian genomes. The results were confirmed on a dataset of orthologous exons with wider species diversity. Thus, the relative contribution of each parameter to the expansion of AARs has been determined.

Our results indicate that AAR expansion is not causally associated to protein adaptive evolution on a genome scale. However, there is a minor contribution of the GC context surrounding the AARs for an increased slippage rate. AARs are over-represented in genes involved in DNA binding and transcriptional activity.

## Results

### Recent expansions in mammalian Positively Selected Genes

Under the hypothesis of AARs as a resource for adaptation, genes that have experienced adaptive evolution are expected to show more and more recent (i.e. purer) AARs associated with a higher substitution rate. To test this prediction, we used the PSGs identified in a thorough study of mammalian genes [17]. First, we compared the amount of repeat containing genes (RCGs) and non-repeat containing genes (non-RCGs) between positively selected genes (PSGs) and non-positively selected genes (non-PSGs) (Table 1). A Fisher's Exact Test shows a weak but significant association between repeats and positive selection ( $p = 0.042$ ). Repeats were then split in two classes, young repeats with high purity ( $\geq 0.9$ ) and old repeats with low purity ( $< 0.9$ ) (Table 1). The PSGs have significantly more young repeats ( $p = 0.0004$ ), suggesting that adaptive evolution in mammals could be associated with recent expansion of repeats.

We also analyzed the physical properties of the AARs. The Lehninger classification describes four categories of

amino acids: acidic, basic, polar uncharged and hydrophobic amino acids [6]. All simple amino acid repeats were classified into the corresponding category for PSGs and non-PSGs (Table 2). The distribution of amino acid repeats differed significantly between PSGs and non-PSGs in a chi-square test ( $p = 0.0003$ ). The differences remain significant after Yate's correction for continuity [18] (Yates'  $p = 0.001$ ) and are mainly due to an excess of repeats of acidic and hydrophobic amino acids in the PSGs. The excess of repeats of hydrophobic AARs explains 77.3% of the differences between PSGs and non-PSGs. However this excess is essentially due to an excess of Leucine repeats. Removing these, the Chi-square is not significant after Yate's correction for continuity (Yates'  $p = 0.067$ ).

### The correlation of amino acid repeats with positive selection and evolutionary rates is spurious

Previous studies in human and mouse have suggested that AAR expansion could be favoured by relaxed purifying selection, repeat length being associated with higher rates of non-synonymous substitutions [13,15]. While our analyses of 6 high-quality mammalian genomes confirm a positive correlation between  $d_N$  and repeat length ( $\rho = 0.043$ ,  $p = 0.002$ ), this is very weak. A stronger correlation is observed between the average purity of AARs and  $d_N$  ( $\rho = 0.111$ ,  $p = 1.54 \cdot 10^{-12}$ ), but there is a similar correlation with  $d_S$  ( $\rho = 0.112$ ,  $p = 7.8 \cdot 10^{-13}$ ), and the correlation with  $\omega$ , which should be most indicative of selection, is the weakest ( $\rho = 0.058$ ,  $p = 0.00017$ ). The similar values of correlation with  $d_N$  and  $d_S$  may be related to the correlations between these rates ( $d_N$  vs.  $d_S$   $\rho = 0.485$ ,  $p < 2.16 \cdot 10^{-16}$ ), and with the GC context surrounding the repeats ( $d_N$  vs.  $GC_{context}$   $\rho = 0.115$ ,  $p < 2.16 \cdot 10^{-16}$ ;  $d_S$  vs.  $GC_{context}$   $\rho = 0.478$ ,  $p < 2.16 \cdot 10^{-16}$ ). Indeed the  $GC_{context}$  also correlates with the purity ( $\rho = 0.09$ ,  $p = 4.272 \cdot 10^{-08}$ ) and the number of AARs ( $\rho = 0.06$ ,  $p < 2.16 \cdot 10^{-16}$ ).

In order to disentangle the effect of these features of gene evolution we fitted the observed variation to a linear model and performed an analysis of variance [e.g. [19]]. We performed this analysis on 3 different mammalian datasets: PSGs, the 6 high-coverage genomes, and orthologous exons (Material and Methods). We detail only the analyses of the PSG dataset (Tables 3 and 4). The other two datasets, with a majority of genes under purifying selection (mean  $\omega = 0.161 \pm 0.21$ ), provide similar results and conclusions with slight variations in the percentage of explained variance (Additional file 1, Tables S1-S4). Adaptive AAR expansions should result in high average purities (i.e., recent or frequent slippage events) and many AARs per positively selected gene. Although the contribution of evolutionary parameters is statistically significant, it is minimal and unlikely to be biologically relevant. For the average purity of the repeats on a gene,  $\omega$  explains only 0.4% of the variance, while the fact of detecting adaptive

**Table 1: Counts of AARs in Positively versus non-Positively Selected Genes in Mammals**

|          | RCGs | non-RCGs | Pure | Impure |
|----------|------|----------|------|--------|
| PSGs     | 19   | 381      | 26   | 8      |
| non-PSGs | 1207 | 14922    | 2021 | 2448   |

Counts of repeat containing genes (RCGs), repeat-free genes (non-RCGs), and of number of pure and impure amino-acid repeats (AARs), of the PSGs and non-PSGs classes. These numbers were used to perform two different Fisher's Exact Tests.

**Table 2: Physicochemical Properties of the AARs in Positively Selected versus Non-Positively Selected Mammalian Genes**

|                 | Acidic       | Basic       | Polar       | Hydrophobic  |
|-----------------|--------------|-------------|-------------|--------------|
| <b>PSGs</b>     | 10 (0.95)    | 0 (-1.08)   | 7 (-2.51)   | 17 (3.23)    |
| <b>non-PSGs</b> | 970 (-0.083) | 154 (0.094) | 2314 (0.22) | 1031 (-0.28) |

Counts of amino acid categories using the Lenhinger classification for each AAR in PSGs and non-PSGs. Values shown in brackets correspond to the residuals for each cell obtained in a Pearson's  $\chi^2$  test.

evolution on any branch of the tree (i.e. significant Likelihood Ratio Test) explains <0.1% of the variance observed for the number of repeats. This shows that the enrichment for recent repeats observed using Fisher's Exact Test was a spurious association. Protein length explains 2% of the variance for AARs, which is not surprising as longer proteins have a greater potential to host repeats. Of note, it has been shown that positive selection tests are also more significant on longer proteins [e.g. [19]], which may contribute to the association between PSGs and AARs.

The excess of leucine repeats also appears spurious, as there is no significant correlation between the  $\omega$  values of each branch in the tree and the length of the leucine repeats ( $\rho = 0.36$ ,  $p = 0.25$ ) or their purity ( $\rho = -0.17$ ,  $p = 0.59$ ).

#### GC rich contexts can favour the expansion of amino acid repeats

The  $GC_{context}$  is the only parameter highly significant in both analyses of variance (on AAR purity and on AAR number). It explains only 1.6% and 0.7% of the variance, but this is 3-fold more than the percentage explained by  $\omega$  or by significant evidence of positive selection. Thus GC-rich sequences appear more prone to the expansion of repeats. To explore this question, we analyzed 16 exons showing accelerated evolution in primates due to GC-biased gene conversion (gBGC) [20]. Two out of these 16 exons have AARs, or 12% of this small dataset. Interestingly, the purity of these repeats highly correlates with the  $GC_{context}$  ( $\rho = 0.85$ ,  $p = 0.002$ , in 10 mammalian

sequences), indicating that a GC increase due to gBGC might sometimes favour the expansion of AARs.

Previous studies have also shown that nucleotide compositional constraints increasing the GC content at 3<sup>rd</sup> codon positions (GC3) influence the expansion of homopolymeric AARs in mammalian and reptilian transcription factors [21]. Analyses of mammalian exons and of complete protein coding genes (Figure 1) shows that there is a weak, but highly significant, positive correlation between purity and GC3 in the DNA sequence surrounding the repeats ( $\rho = 0.28$ ,  $p < 2.2 \cdot 10^{-16}$  and  $\rho = 0.126$ ,  $p < 2.2 \cdot 10^{-16}$ , for exons and whole genes, respectively). A Welch's t-test comparing the GC3 context of exons containing pure and impure repeats indicates that genes hosting pure repeats have on average a higher GC3 than impure repeats (0.75 and 0.66 respectively,  $p < 2.2 \cdot 10^{-16}$ ). In summary, these results consistently indicated that in mammals there is a small but significant increase of AAR expansion in regions with high GC.

#### Amino acid repeats and gene expression

The main reasons that led us to study the relationship between repeat expansion and expression levels are: 1) The observed excess of hydrophobic repeats is likely to lead to aggregation and misfolding in PSGs [22]. 2) The correlation between substitution rates and  $GC_{context}$  that also correlates with the average purity of AARs, has been shown to be limited by expression-related purifying selection [23]. 3) In *E. coli* it has been observed that the stability of the structure around the translation start is directly related with the expression level [24].

**Table 3: ANOVA of Linear Model to Explain the Average Purity of the AARs in Positively Selected and Non-Positively Selected Genes**

|                             | Df   | Sum Sq  | Mean Sq | F value | p-value         | Var. (%) <sup>6</sup> |
|-----------------------------|------|---------|---------|---------|-----------------|-----------------------|
| Residuals                   | 3616 | 20.5105 | 0.0057  |         |                 | 97.351                |
| $GC_{context}$ <sup>1</sup> | 1    | 0.3351  | 0.3351  | 59.078  | <b>1.94E-14</b> | <b>1.590</b>          |
| Species <sup>2</sup>        | 5    | 0.1154  | 0.0231  | 4.0684  | <b>0.001101</b> | <b>0.548</b>          |
| $\omega$ <sup>3</sup>       | 1    | 0.0872  | 0.0872  | 15.3805 | <b>8.95E-05</b> | <b>0.414</b>          |
| LRT <sup>4</sup>            | 1    | 0.0183  | 0.0183  | 3.2305  | 0.072362        | 0.087                 |
| P. length (aa) <sup>5</sup> | 1    | 0.002   | 0.002   | 0.3525  | 0.552754        | 0.009                 |
| Total                       | 3625 | 21.0685 | 0.4714  |         |                 |                       |

<sup>1</sup>GC content excluding the stretch containing AARs; <sup>2</sup>species containing the AAR(s); <sup>3</sup>omega ( $d_N/d_S$ ) of the most significant evolutionary model;

<sup>4</sup>significant test for positive selection at any branch of the tree; <sup>5</sup>protein length in aminoacids; <sup>6</sup>proportion of variance explained.

**Table 4: ANOVA of Linear Model to Explain the Number of AARs in Positively Selected and Non-Positively Selected Mammalian Genes**

|                             | Df    | Sum Sq  | Mean Sq | F value | p-value             | Var. (%) <sup>6</sup> |
|-----------------------------|-------|---------|---------|---------|---------------------|-----------------------|
| Residuals                   | 82096 | 6806.8  | 0.1     |         |                     | 96.879                |
| P. length (aa) <sup>1</sup> | 1     | 168.1   | 168.1   | 2027.45 | <b>&gt;2.20E-16</b> | <b>2.392</b>          |
| GCcontext <sup>2</sup>      | 1     | 48.9    | 48.9    | 590.12  | <b>&gt;2.20E-16</b> | <b>0.696</b>          |
| LRT <sup>3</sup>            | 1     | 1.4     | 1.4     | 16.3141 | <b>&gt;3.71E-06</b> | <b>0.020</b>          |
| Species <sup>4</sup>        | 5     | 0.8     | 0.2     | 1.9078  | 0.0894              | 0.011                 |
| $\omega$ <sup>5</sup>       | 1     | 0.048   | 0.04798 | 0.5787  | 0.4468              | 0.001                 |
| Total                       | 82105 | 7026.01 | 218.748 |         |                     |                       |

<sup>1</sup>Protein Length in aminoacids; <sup>2</sup>GC content excluding the stretch containing AARs; <sup>3</sup>significant test for positive selection at any branch of the tree; <sup>4</sup>species containing the AAR(s); <sup>5</sup> $d_N/d_S$  of the most significant evolutionary model; <sup>6</sup>proportion of variance explained.

For the 1,057 human and 1,009 mouse genes that contain at least one AAR, we performed an analysis of variance including the expression levels in 5 representative organs as factors. The result shows that expression level has no impact on the expansion of AARs, measured as average purity or as number of repeats in the hosting gene (Additional file 1, Tables S5-S8), neither in mouse nor human.

Conversely, the number of AARs proximal to the translation start for human and mouse does not explain, in any of the 5 organs, the observed variance in the expression levels. For simplicity we show only the results obtained for the human brain (Table 5).

In conclusion, we can reject any simple relation between the presence of AARs or their age, and the expression level of human and mouse genes.

#### Molecular function of genes hosting amino acid repeats

We studied the relation between AARs and the Gene Ontology terms (GO), for Molecular Function, Biological Process and Cell Component, of all human and mouse protein-coding genes. As very similar results were obtained for both species we will report only those obtained for human.

Genes containing AARs are enriched in a wide variety of molecular functions, mainly involved in binding, transcription and nuclear structures (Table 6); analyses accounting for purity or Biological Process of genes with AARs support these results (data not shown). Including these molecular function terms in the linear model to explain the number of AARs per gene, the total percentage of variance explained by significantly enriched GO terms is 13.9% for human and 15.2% for mouse (see Table 7 for human and Table S9 for mouse). This is not the case for average purity of AARs, for which GC context remains the main explanatory factor in human (2.73% of variance explained, Table S10). Finally, the cellular compartment

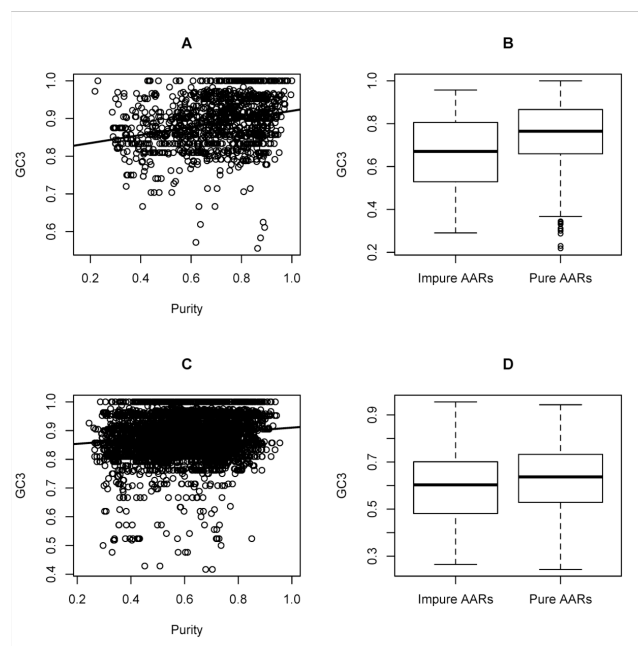
nucleus is also enriched in genes with AARs, and in genes with purer AARs (GO:0005634,  $p < 6.19 \cdot 10^{-12}$ ).

The ice binding molecular function (GO:0050825) is overrepresented. But this excess disappears after excluding the Alanine repeats. This appears to be an annotation bias, as genes containing alanine-rich repeats are attributed this function by partial sequence similarity with the InterPro entry IPR000104 (Antifreeze protein, type I), a special glycoprotein identified in marine teleosts from polar oceans[25].

#### Discussion

In mammals, a positive correlation between  $d_N$  and repeat length is weak but statistically significant. This result is congruent with previous analyses in smaller datasets of human and mouse genomes [13,15]. The purity of the AARs per gene or exon shows a similar trend. But these weak correlations can be explained by the influence of the GC context surrounding the repeat. High GC content can generate a sequence context more prone to slip-page[21,26-28] and thus expansion of AARs. Indeed we found an example of this in exons that have experienced GC-biased gene conversion in primates. Similarly, while there is an increase in the amount of recent AARs in mammalian PSGs, these recent expansions are better explained by GC content than by positive selection acting on codons. Therefore it seems that, in contradiction to previous reports [15], the expansion of AARs is not causally associated with substitution rates. While purifying selection limits the expansion of AARs[e.g. [29]], this appears to be distinct from the selective pressure on individual (aligned) amino acid sites. That means that these repeats are experiencing not only different mutational processes, but also particular selective constraints, leading to a more complex scenario of evolution.

Our analyses, even of individual exons, suggest that increased substitution rates are not usually linked to the



**Figure 1** of GC content at 3<sup>rd</sup> codon position on AAR purity. **Influence of GC content at 3<sup>rd</sup> codon position on AAR purity.** GC3, GC at 3<sup>rd</sup> codon positions in the sequence context of the repeats. (A) positive correlation and regression line (using least squares) between GC3 and purity in orthologous mammalian exons; (B) Average GC3 in Impure and Pure AARs in orthologous mammalian exons ( $p < 2.16 \cdot 10^{-16}$ ; Welch's t-test); (C) positive correlation between GC3 and purity in mammalian genomes and regression line (using least squares); (D) Average GC3 in Impure and Pure AARs in mammalian genomes ( $p < 2.16 \cdot 10^{-16}$ ; Welch's t-test).

presence of AARs. However, it is possible that in some particular cases, as has been suggested for *Drosophila*, the expansion of AARs can produce compensatory changes on the neighbouring sites to accommodate the perturbation generated by the repeat[30]. We also cannot exclude the existence of adaptive evolution related with AARs[7,8], in the absence of a good reference neutral model for trinucleotide expansions in proteins. But our results do show that the selective pressure as measured by codon models is not related with putative adaptive evolution of AARs.

AARs in mammalian genes do not seem to affect gene expression significantly. Unlike repeats which disrupt the reading frame, and have a strong effect on replication and transcription stability[31], the tri-nucleotide repeats might be constrained in a different way. It seems that repeats located in the promoter region[32] have a stronger

influence on transcription than do AARs, even those near the transcription start.

The analyses of molecular function confirmed an enrichment in the transcription factor, DNA binding, molecular transducers and binding categories that is consistent with previous studies of polymorphic repeats [26,33,34]. The overrepresentation of transcription factor categories supports the existence of *trans* effects, as these repeats might alter the expression of the target genes and end up producing dramatic changes on the phenotype[7]. However, while the ice-binding protein is involved in hypothermic resistance in some antarctic fishes vertebrates[25,35], its overrepresentation in alanine-rich mammalian genes is probably due to an annotation bias.

In general, we found that AARs are located in proteins that interact with DNA, RNA, ligands or other proteins, so it is likely that they contribute to adapt or modulate the inter-

**Table 5: ANOVA of a Linear Model to Explain the Expression Level of Human Genes in the Brain**

|                             | Df  | Sum Sq | Mean Sq | F value | p-value |
|-----------------------------|-----|--------|---------|---------|---------|
| P. length (aa) <sup>1</sup> | 1   | 2.5    | 2.5     | 0.6648  | 0.4151  |
| GCcontent <sup>2</sup>      | 1   | 0.1    | 0.1     | 0.0178  | 0.894   |
| N° AARs <sup>3</sup>        | 1   | 0.1    | 0.1     | 0.0226  | 0.8805  |
| AARs +30 nt <sup>4</sup>    | 1   | 1      | 1       | 0.2669  | 0.6055  |
| AARs +60 nt <sup>5</sup>    | 1   | 1.3    | 1.3     | 0.3386  | 0.5608  |
| AARs +90 nt <sup>6</sup>    | 1   | 5.5    | 5.5     | 1.4469  | 0.2293  |
| d <sub>N</sub> <sup>7</sup> | 1   | 10.1   | 10.1    | 2.6413  | 0.1045  |
| Average Purity <sup>8</sup> | 1   | 0.4    | 0.4     | 0.114   | 0.7357  |
| Residuals                   | 893 | 3416.8 | 3.8     |         |         |

<sup>1</sup>GC content excluding the stretch containing AARs; <sup>2</sup>protein length in aminoacids; <sup>3</sup>Number of AARs; <sup>4-6</sup>Number of AARs in a window of +30 nt, +60 nt and +90 nt from translation start; <sup>7</sup>Non-synonymous substitution rate; <sup>8</sup>Average Purity of the AARs.

action capacity of these proteins. Longer proteins and repeat-rich proteins tend to have a higher connectedness within interaction networks, suggesting that they contribute to an enlarged interaction surface and constitute more flexible subunits[36]. Some AAR have been recently associated to the presence of repeats to specific domains, such as signal peptides or transmembrane regions[16], pointing to their role in facilitating molecular interactions of extreme importance. For example, in the *Drosophila* ARC 70 cofactor complex, the -130 and -230 subunits contain an expansion of glutamine residues, a prevalent feature of sequence-specific activators in *Drosophila*[37].

## Conclusions

Despite the appealing idea of an adaptive role of the expansion of amino acid repeats, we can rule out a link with adaptive evolution in mammalian protein-coding genes as measured by codon models. Genome-wide, GC content is more relevant to amino acid repeat expansions than substitution rates. Amino acid repeats are under strong functional constraints and expand preferentially in transcription factors and nuclear genes involved in DNA and/or protein interactions. Why some genes accumulate more and most recent amino acid repeats requires further study in a network context, to shed light on the evolutionary dynamics and function of these mutations.

## Methods

### Positively Selected Genes (PSGs)

A recent study in mammals[17] performed a thorough analysis for detecting positive selection in six mammalian genomes. A likelihood ratio test for positive selection on any branch of the phylogeny reported 400 Positively Selected Genes (PSGs), and 16,129 genes that have not experienced any detected positive selection in mammals (non-PSGs). Alignments for these genes were downloaded from the author's website <http://comp.gen.bscb.cornell.edu/projects/mammal-psg/lrtall.txt> and screened for repeats.

### High-quality Mammalian Genomes

To study the relationship of multiple factors that could be influencing the expansion of repeats in mammalian genomes, we used mammalian assemblies with high cov-

**Table 6: Enrichment of Molecular Functions of Genes containing AARs**

| GO.ID      | Term <sup>1</sup>                                                 | Corrected p-value <sup>2</sup> |
|------------|-------------------------------------------------------------------|--------------------------------|
| GO:0050825 | <b>ice binding</b>                                                | < 1E-26                        |
| GO:0003677 | <b>DNA binding</b>                                                | 4.01E-15                       |
| GO:0003700 | <b>transcription factor activity</b>                              | 1.26E-13                       |
| GO:0043565 | <b>sequence-specific DNA binding</b>                              | 5.79E-13                       |
| GO:0005199 | structural constituent of cell wall                               | 1.00E-08                       |
| GO:0004879 | <b>ligand-dependent nuclear receptor activity</b>                 | 3.15E-07                       |
| GO:0003682 | <b>chromatin binding</b>                                          | 2.54E-06                       |
| GO:0003723 | RNA binding                                                       | 7.63E-05                       |
| GO:0008270 | zinc ion binding                                                  | 0.000303826                    |
| GO:0004969 | histamine receptor activity                                       | 0.0008013                      |
| GO:0045735 | nutrient reservoir activity                                       | 0.0008013                      |
| GO:0003702 | RNA polymerase II transcription factor activity                   | 0.001116964                    |
| GO:0003676 | nucleic acid binding                                              | 0.001580342                    |
| GO:0003705 | RNA polymerase II transcription factor activity, enhancer binding | 0.009862154                    |
| GO:0003735 | structural constituent of ribosome                                | 0.02671                        |
| GO:0005249 | voltage-gated potassium channel activity                          | 0.049858667                    |
| GO:0004386 | helicase activity                                                 | 0.065105625                    |
| GO:0016563 | transcription activator activity                                  | 0.13355                        |
| GO:0003714 | transcription corepressor activity                                | 0.13355                        |
| GO:0005179 | hormone activity                                                  | 0.199622105                    |

<sup>1</sup> In bold terms overrepresented also for genes hosting the highest average purity of their AARs; <sup>2</sup> FDR < 20%.

**Table 7: Percentage of Explained Variance of the Number of Aminoacid Repeats**

| Factor                                                            | Pr(>F)    | Var. (%)    |
|-------------------------------------------------------------------|-----------|-------------|
| <b>ice binding</b>                                                | <2.20E-16 | 5.869336006 |
| P. length                                                         | <2.20E-16 | 2.718369933 |
| structural constituent of cell wall                               | <2.20E-16 | 1.965991088 |
| <b>DNA binding</b>                                                | <2.20E-16 | 1.544242393 |
| GC context                                                        | <2.20E-16 | 0.754548334 |
| structural constituent of ribosome                                | <2.20E-16 | 0.597911216 |
| <b>Transcription factor activity</b>                              | <2.20E-16 | 0.575348528 |
| hormone activity                                                  | <2.20E-16 | 0.554521432 |
| histamine receptor activity                                       | <2.20E-16 | 0.553219739 |
| nucleic acid binding                                              | <2.20E-16 | 0.547145169 |
| Voltage-gated potassium channel activity                          | <2.20E-16 | 0.488135064 |
| <b>ligand-dependent nuclear receptor activity</b>                 | 2.33E-12  | 0.348853859 |
| <b>sequence-specific DNA binding</b>                              | 3.01E-09  | 0.249491255 |
| RNA binding                                                       | 1.70E-07  | 0.193952332 |
| $d_S$                                                             | 1.25E-06  | 0.166616768 |
| <b>chromatin binding</b>                                          | 3.29E-06  | 0.153165936 |
| RNA polymerase II transcription factor activity, enhancer binding | 3.63E-06  | 0.151864242 |
| $d_N$                                                             | 6.19E-06  | 0.144921877 |
| nutrient reservoir activity                                       | 0.0004664 | 0.086779567 |
| transcription corepressor activity                                | 0.0054142 | 0.054671127 |
| $\omega$                                                          | 0.0134962 | 0.043389783 |
| RNA polymerase II transcription factor activity                   | 0.0240022 | 0.03601352  |
| helicase activity                                                 | 0.1667501 | 0.013450833 |
| zinc ion binding                                                  | 0.198911  | 0.011715242 |
| transcription activator activity                                  | 0.4614908 | 0.003905081 |

In italics GO Terms that remain significant after Bonferroni Correction. In Bold functions enriched in pure AARs.

erage (ranging from 6-11×) and their corresponding Ensembl 50 Genes[38]. We compared the genomes of 2 primates (*Homo sapiens* NCBI36 and *Pan troglodytes* CHIMP2.1), 2 rodents (*Mus musculus* NCBI37 and *Rattus norvegicus* RGC3.4) and 2 domestic species (*Bos taurus* Btau\_3.1 and *Canis familiaris* Canfam 2.0).

For each mammalian genome, we downloaded all the known protein coding genes, with exception of dog and chimp genomes where, in order to gather the largest accurate dataset, we used the "known by projection" set. The repeat analyses are restricted to non-redundant one-to-one orthologues to an equidistant outgroup, dog in the case of rodents and primates, and human for the domestic species. We filtered the genes by keeping the protein corresponding to the longest transcript and excluding all coding sequences that did not begin with a start codon. Finally the number of genes that were screened for repeats in each species was 13,926 human, 11,120 chimpanzee, 13,921 mouse, 10,360 rat, 7,073 cow and 7,834 dog genes.

#### Orthologous Exons

We downloaded 1,168 orthologous exons alignments including 9 to12 mammalian species, from the OrthoMam database [39]. This is a curated database that contains the amino acid and coding sequence alignments for each particular exon. The inclusion of these align-

ments allowed studying local AAR expansions without biases due to regional differences in substitution rates and GC context along the whole gene. The exon trees were built using PHYML (substitution model = JTT, estimated proportion of invariable sites, four categories, estimated gamma, initial tree with BIONJ) [40]. Evolutionary rates for each branch were obtained running the *free-ratios* model in PAML 4.1 [41] and keeping  $d_N$ ,  $d_S$  and  $\omega$  convergent values of 5 replicate runs. Non-convergent or 999 values were not considered in further analyses.

#### Homo-polymeric Amino-acid Repeats and Purity

As in many previous studies we focused on perfect homo-polymeric amino-acid repeats, where we assume that the expansion of a tri-nucleotide by slippage gave birth to the repetition of a single amino-acid motif within the protein. To consider that an amino-acid repeat appeared by polymerase slippage a minimum threshold of 5 units was frequently used in the literature [e.g. [8,26]]. We used a minimum number of 7 units. The reasons for this are, first, to increase the significance level[6] and, second, to increase the chance that a repeat locus shows length polymorphism [42,43].

The *purity* of the nucleotide sequence coding the amino-acid repeat was calculated following the method described by Laidlaw et al. in 2007[8] that is summarized in the equation below;

$$Purity = \frac{(m-n)}{m} \quad (1)$$

where  $m$  is the total number of nucleotides coding the amino-acid repeat and  $n$  the number of interruptions or nucleotide changes with respect to the canonical codon (the most frequent or most likely to have experienced expansion by slippage). The presence of AARs was considered for each species independently of the presence of that repeat in orthologues.

#### Summary of parameters and estimates

Each gene was screened for homo-polymeric amino acid repeats within the corresponding protein sequence. The following parameters were calculated:

i) *Weighted Average Purity of the Repeats of a Gene*: the weighted average of the *purity* estimates of every amino-acid repeat in the protein sequence of a gene. The weighting is based on the length of the coding sequence of each individual repeat, as described in the following equation;

$$Av.Purity = \frac{\sum_i l_i \cdot P_i}{L} \quad (2)$$

where  $n$  is the total number of AARs on the protein-coding gene,  $l$  is the length in bp of each individual repeat,  $P$  is the corresponding Purity and  $L$  the sum in bp of the length of all AARs in the gene. This measure allowed us to compare if certain genes contain purer AARs than others. Note that the vast majority of the cases correspond to genes hosting only one AAR. (Additional file 2, Figures SF1 and SF2)

ii)  $d_N$  and  $d_S$ : sitewise maximum likelihood estimates of  $d_N$  and  $d_S$  for each orthologous pair were downloaded from Ensembl [38].

iii) *GC context* (%GC): the GC content of gene after excluding all regions encoding repeats[44]. Similarly, we estimated GC3 as the GC content on third codon positions of the full repeat-free coding dna of the gene or exon. These parameters depict the sequence context in which repeats are born.

#### Gene Expression data

Microarray data of mouse and human tissues were downloaded from ArrayExpress (E-AFMX-4 and E-AFMX-5)[45]. E-AFMX-4 uses an Affymetrix Custom Array - Novartis Mouse (A-AFFY-39) and E-AFMX-5 uses an Affymetrix Custom Array - Novartis Human (A-AFFY-40). Mapping of Ensembl gene on Affymetrix probesets from these chips was taken from <http://biogps.gnf.org/downloads/>. E-AFMX-5 also uses an Affymetrix GeneChip

Human Genome HG-U133A (A-AFFY-33), whose mapping to Ensembl genes was downloaded from BioMart [46].

We extracted expression data for 5 organs in mouse (cerebral cortex, liver, kidney, testis and heart) and human (brain, liver, kidney, testis and heart). Raw CEL files were renormalized using the package gcRMA [47] of Bioconductor version 2.2[48]. We used the "affinities" model of gcRMA, which uses mismatch probes as negative control probes to estimate the non-specific binding of probe sequences. The normalized values of expression are in log2 scale, which attenuates the effect of outliers. Expression values were averaged between replicates and between multiple probes mapped to a same gene. Probes mapping to more than one gene were discarded.

#### GO term enrichment

Over and under representation of GO terms [49] was tested by means of a Fisher exact test, using the Bioconductor package topGO version 1.8.1 [50]. The reference set was all Ensembl genes used in the repeats analysis. The GO annotation of Ensembl genes was downloaded from BioMart. The "elim" algorithm of topGO was used, allowing to decorrelate the graph structure of the gene ontology, reducing non-independence problems. Gene ontology categories with a FDR < 20% were reported.

#### Authors' contributions

FC designed the project, gathered the genomic data, performed most of the evolutionary and statistical analyses, and wrote the original manuscript. JR gathered the expression and gene ontology data, and performed the analysis of GO Term enrichment. MRR supervised the work, provided critical comments about the statistical analyses and the biological discussion, and revised thoroughly the manuscript. All authors have read and approved the final manuscript.

#### Additional material

##### Additional file 1

**Supplementary Tables.** A PDF file containing additional Tables S1-S11. These tables contain analyses of variance with factors sorted by percentage of explained variance. Further details are provided as footnotes accompanying each table.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-10-619-S1.PDF>]

##### Additional file 2

**A PDF file containing additional Figures SF1-SF5.** Further details are provided as footnotes accompanying each Figure.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2164-10-619-S2.PDF>]

## Acknowledgements

We acknowledge funding from Etat de Vaud, Swiss National Science Foundation grant I16798, and the Swiss Institute for Bioinformatics. We would like to thank Nicolas Galtier and Emmanuel Douzery for helping with the exon alignments. We also thank Carolin Kosiol, Nicolas Salamin, Matthew T. Webster, Carles Vilà and anonymous referees for their helpful comments.

## References

- Ellegren H: **Microsatellite mutations in the germline: implications for evolutionary inference.** *Trends in Genetics* 2000, **16**(12):551.
- Kashi Y, King DG: **Simple sequence repeats as advantageous mutators in evolution.** *Trends in Genetics* 2006, **22**(5):253-259.
- Kruglyak S, Durrett RT, Schug MD, Aquadro CF: **Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations.** *Proceedings of the National Academy of Sciences of the United States of America* 1998, **95**(18):10774-10778.
- Ellegren H: **Microsatellites: simple sequences with complex evolution.** *Nat Rev Genet* 2004, **5**(6):435-445.
- Beena T, Koshy HYZ: **The CAG/Polyglutamine Tract Diseases: Gene Products and Molecular Pathogenesis.** *Brain Pathology* 1997, **7**(3):927-942.
- Karlin S, Brocchieri L, Bergman A, Mrázek J, Gentles AJ: **Amino acid runs in eukaryotic proteomes and disease associations.** *Proceedings of the National Academy of Sciences of the United States of America* 2002, **99**(1):333-338.
- Fondon JW III, Garner HR: **Molecular origins of rapid and continuous morphological evolution.** *Proceedings of the National Academy of Sciences* 2004, **101**(52):18058-18063.
- Laidlaw J, Gelfand Y, Ng K-W, Garner HR, Ranganathan R, Benson G, Fondon JW III: **Elevated Basal Slippage Mutation Rates among the Canidae.** *J Hered* 2007, **98**(5):452-460.
- Zamorzeva I, Rashkovetsky E, Nevo E, Korol A: **Sequence polymorphism of candidate behavioural genes in *Drosophila melanogaster* flies from 'Evolution canyon'.** *Molecular Ecology* 2005, **14**(10):3235-3245.
- Nevo E, Beharav A, Meyer RC, Hackett CA, Forster BP, Russell JR, Powell W: **Genomic microsatellite adaptive divergence of wild barley by microclimatic stress in 'Evolution Canyon', Israel.** *Biological Journal of the Linnean Society* 2005, **84**(2):205-224.
- Fahima T, Röder MS, Wendehake K, Kirzhner VM, Nevo E: **Microsatellite polymorphism in natural populations of wild emmer wheat, *Triticum dicoccoides*, in Israel.** *TAG Theoretical and Applied Genetics* 2002, **104**(1):17-29.
- Hammock EAD, Young LJ: **Microsatellite Instability Generates Diversity in Brain and Sociobehavioral Traits.** *Science* 2005, **308**(5728):1630-1634.
- Mularoni L, Veritia RA, Albà MM: **Highly constrained proteins contain an unexpectedly large number of amino acid tandem repeats.** *Genomics* 2007, **89**(3):316-325.
- Albà MM, Santibanez-Koref MF, Hancock JM: **Conservation of polyglutamine tract size between mice and humans depends on codon interruption.** *Mol Biol Evol* 1999, **16**(11):1641-1644.
- Hancock JM, Worthey EA, Santibanez-Koref MF: **A Role for Selection in Regulating the Evolutionary Emergence of Disease-Causing and Other Coding CAG Repeats in Humans and Mice.** *Mol Biol Evol* 2001, **18**(6):1014-1023.
- Simon M, Hancock J: **Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins.** *Genome Biology* 2009, **10**(6):R59.
- Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A: **Patterns of Positive Selection in Six Mammalian Genomes.** *PLoS Genetics* 2008, **4**(8):e1000144.
- Preacher KJ: **Calculation for the chi-square test: An interactive calculation tool for chi-square tests of goodness of fit and independence [Computer software].** [<http://www.quantpsy.org>].
- Studer RA, Penel S, Duret L, Robinson-Rechavi M: **Pervasive positive selection on duplicated and nonduplicated vertebrate protein coding genes.** *Genome Research* 2008, **18**(9):1393-1402.
- Galtier N, Duret L, Glémin S, Ranwez V: **GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates.** *Trends in Genetics* 2009, **25**(1):1-5.
- Nakachi Y, Hayakawa T, Oota H, Sumiyama K, Wang L, Ueda S: **Nucleotide compositional constraints on genomes generate alanine-, glycine-, and proline-rich structures in transcription factors.** *Mol Biol Evol* 1997, **14**(10):1042-1049.
- Oma Y, Kino Y, Toriumi K, Sasagawa N, Ishiura S: **Interactions between homopolymeric amino acids (HPAAs).** *Protein Science* 2007, **16**(10):2195-2204.
- Drummond DA, Wilke CO: **Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution.** 2008, **134**(2):341-352.
- Kudla G, Murray AW, Tollervey D, Plotkin JB: **Coding-Sequence Determinants of Gene Expression in *Escherichia coli*.** *Science* 2009, **324**(5924):255-258.
- Sicheri F, Yang DSC: **Ice-binding structure and mechanism of an antifreeze protein from winter flounder.** *Nature* 1995, **375**(6530):427-431.
- Mularoni L, Guigo R, Albà MM: **Mutation patterns of amino acid tandem repeats in the human proteome.** *Genome Biology* 2006, **7**(4):R33.
- Brock GJ, Anderson NH, Monckton DG: **Cis-acting modifiers of expanded CAG/CTG triplet repeat expandability: associations with flanking GC content and proximity to CpG islands.** *Hum Mol Genet* 1999, **8**(6):1061-1067.
- Jurka J, Pethiyagoda C: **Simple repetitive DNA sequences from primates: Compilation and analysis.** *Journal of Molecular Evolution* 1995, **40**(2):120-126.
- Loire E, Praz F, Higuier D, Netter P, Achaz G: **Hypermutability of genes in *Homo sapiens* due to the hosting of long mono-SSR.** *Mol Biol Evol* 2008:111-121.
- Huntley MA, Clark AG: **Evolutionary Analysis of Amino Acid Repeats across the Genomes of 12 *Drosophila* Species.** *Mol Biol Evol* 2007, **24**(12):2598-2609.
- Ackermann M, Chao L: **DNA Sequences Shaped by Selection for Stability.** *PLoS Genet* 2006, **2**(2):e22.
- Riley DE, Jeon JS, Krieger JN: **Simple repeat evolution includes dramatic primary sequence changes that conserve folding potential.** *Biochemical and Biophysical Research Communications* 2007, **355**(3):619-625.
- Legendre M, Pochet N, Pak T, Verstrepen KJ: **Sequence-based estimation of minisatellite and microsatellite repeat variability.** *Genome Research* 2007, **17**(12):1787-1796.
- O'Dushlaine C, Edwards R, Park S, Shields D: **Tandem repeat copy-number variation in protein-coding regions of human genes.** *Genome Biology* 2005, **6**(8):R69.
- Hirano Y, Nishimiya Y, Matsumoto S, Matsushita M, Todo S, Miura A, Komatsu Y, Tsuda S: **Hypothermic preservation effect on mammalian cells of type III antifreeze proteins from notched-fin eelpout.** *Cryobiology* 2008, **57**(1):46-51.
- Dosztanyi Z, Chen J, Dunker AK, Simon I, Tompa P: **Disorder and Sequence Repeats in Hub Proteins and Their Implications for Network Evolution.** *Journal of Proteome Research* 2006, **5**(11):2985-2995.
- Levine M, Tjian R: **Transcription regulation and animal diversity.** *Nature* 2003, **424**(6945):147-151.
- Hubbard TJP, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, et al.: **Ensembl 2009.** *Nucleic Acids Research* 2009, **37**(suppl\_1):D690-697.
- Ranwez V, Delsuc F, Ranwez S, Belkhir K, Tilak M-K, Douzery E: **OrthoMaM: A database of orthologous genomic markers for placental mammal phylogenetics.** *BMC Evolutionary Biology* 2007, **7**(1):241.
- Guindon Sp, Gascuel O: **A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood.** *Systematic Biology* 2003, **52**(5):696-704.
- Yang Z: **PAML 4: Phylogenetic Analysis by Maximum Likelihood.** *Mol Biol Evol* 2007, **24**(8):1586-1591.
- Pupko T, Graur D: **Evolution of Microsatellites in the Yeast *Saccharomyces cerevisiae*: Role of Length and Number of Repeated Units.** *Journal of Molecular Evolution* 1999, **48**(3):313-316.
- Weber JL: **Informativeness of human (dC-dA)<sub>n</sub>-(dG-dT)<sub>n</sub> polymorphisms.** *Genomics* 1990, **7**(4):524-530.



44. Albà MM, Guigo R: **Comparative Analysis of Amino Acid Repeats in Rodents and Humans.** *Genome Res* 2004, **14(4)**:549-554.
45. Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, Abeygunawardena N, Berube H, Dylag M, Emam I, Farne A, et al.: **ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression.** *Nucl Acids Res* 2009, **37(suppl\_1)**:D868-872.
46. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, Kasprzyk A: **BioMart - biological queries made easy.** *BMC Genomics* 2009, **10(1)**:22.
47. Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F: **A Model-Based Background Adjustment for Oligonucleotide Expression Arrays.** *Journal of the American Statistical Association* 2004, **99(468)**:909-917.
48. Gentleman R, Carey V, Bates D, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, et al.: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biology* 2004, **5(10)**:R80.
49. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al.: **Gene Ontology: tool for the unification of biology.** *Nat Genet* 2000, **25(1)**:25-29.
50. Alexa A, Rahnenfuhrer J, Lengauer T: **Improved scoring of functional groups from gene expression data by decorrelating GO graph structure.** *Bioinformatics* 2006, **22(13)**:1600-1607.